THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Experimental and computational exploration of enzyme sequence space

ELZBIETA REMBEZA

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Biology and Biological Engineering
Chalmers University of Technology
Gothenburg, Sweden 2021

Experimental and computational exploration of enzyme sequence space
ELZBIETA REMBEZA

ISBN 978-91-7905-576-9

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie 5043
ISSN 0346-718X

Division of Systems and Synthetic Biology
Department of Biology and Biological Engineering
Chalmers University of Technology
SE 412 96 Gothenburg
Sweden
Telephone +46 (0)31 772 1000

Cover: Representation of sequence space of S-2-hydroxyacid oxidases.

Printed by Chalmers Reproservice
Gothenburg, Sweden 2021

Experimental and computational exploration of enzyme sequence space

ELZBIETA REMBEZA

Department of Biology and Biological Engineering

Chalmers University of Technology

ABSTRACT

Millions of enzymes with desirable features or new exciting activities can be found in organisms occupying diverse niches all around the earth. However, enzyme studies tend to be biased towards characterisation of representatives from eukaryotes, model organisms, or disease-causing bacteria. As such, a large number of enzymes still remains underexplored. The so-called sequence space of proteins - all possible protein sequences - is even greater when we include not only natural sequences, but also the ones designed by human or artificial intelligence. This thesis explores various reasons, approaches, and outcomes of investigation of large enzymatic sequence spaces.

In the first part of my work, I focused on investigation of a natural sequence space of oxidases using a high-throughput activity profiling platform. A functional screen of an industrially important class of enzymes, S-2-hydroxyacid oxidases (EC 1.1.3.15), revealed that nearly 80% of the class is misannotated. Further exploration of annotations to public databases indicated that similar errors of annotations can be found in other enzyme classes. A broader activity profiling of 1.1.3.x oxidases resulted in the discovery of two novel microbial enzymes: N-acetyl-hexosamine oxidase, and a novel type of long-chain alcohol oxidase.

Natural enzymes often need to be improved in order to be industrially applied, for example to become more stable, or accept non-natural substrates. A novel, and constantly developing, approach for enzyme design involves the use of machine learning (ML) tools. Second part of my work focused on screening an enzyme sequence space designed by generative adversarial networks. Our work proved that ML methods can generate fully functional enzymes that mimic sequences present in nature.

Enzyme assays are necessary to get a full understanding of how enzymes work. Traditional kinetic assays are time- and reagent-consuming and as a result a limited number of variants and conditions are being tested for each target. In my final work I described a novel approach for enzyme kinetic studies, by adaptation of a microfluidic qPCR device.

Keywords: enzyme sequence space, high-throughput-screening, enzyme discovery, oxidases, protein annotation

LIST OF PUBLICATIONS

This thesis is based on the work contained in the following publications:

Paper I: Rembeza E, Engqvist MKM. Experimental and computational investigation of enzyme functional annotations uncovers misannotation in the EC 1.1.3.15 enzyme class. *PLoS Comput Biol.* 2021;17: e1009446.

Paper II: Rembeza E, Boverio A, Fraaije MW, Engqvist MKM. Discovery of two novel oxidases using a high-throughput activity screen. *Chembiochem*. 2021. doi:10.1002/cbic.202100510

Paper III: Repecka D, Jauniskis V, Karpus L, Rembeza E, Rokaitis I, Zrimec J, Poviloniene S, Laurynenas A, Viknander S, Abuajwa W, Savolainen O, Meskys R, Engqvist MKM and Zelezniak A. Expanding functional protein sequence spaces using generative adversarial networks. *Nature Machine Intelligence*. 2021;3: 324–333.

Paper IV: Rembeza E, Engqvist MKM. Adaptation of a Microfluidic qPCR System for Enzyme Kinetic Studies. *ACS Omega*. 2021;6: 1985–1990.

CONTRIBUTION SUMMARY

Paper I: I co-designed the study, carried out the experiments, co-analysed the data, wrote the draft, and co-wrote the manuscript.

Paper II: I co-designed the study, carried out majority of the experiments, co-analysed the data, wrote the draft, and co-wrote the manuscript.

Paper III: I performed part of the experiments and co-analysed the data.

Paper IV: I co-designed the study, carried out the experiments, co-analysed the data, wrote the draft, and co-wrote the manuscript.

PREFACE

This dissertation serves as partial fulfilment of the requirements to obtain the degree of Doctor of Philosophy at the Department of Biology and Biological Engineering at Chalmers University of Technology. The PhD studies were carried out between October 2017 and January 2022 at the Division of Systems and Synthetic Biology under the supervision of Martin Engqvist. The research was funded by Chalmers Foundation.

TABLE OF CONTENTS

When we try to pick out anything by itself,
we find it hitched to everything else in the universe.

John Muir

Chapter 1. Introduction

1.1. History of enzyme studies

Enzymes are molecules which facilitate reactions happening in all living organisms. Thanks to them we are able to eat, breathe, move, and think. Well-functioning enzymes ensure a well working organism, while their malfunction might have dire consequences for health [1]. Different organisms contain a different set of enzymes, some of which are crucial for survival in extreme environments or fighting and infecting other organisms [2–4]. Enzymes are also used in our everyday life in detergents, cosmetics, as pharmaceuticals and food additives [5,6]. In fact, people have been relying on enzymatic reactions for commercial purposes for centuries: to ferment sugars into alcohol, produce cheese or leather [7,8]. For many years, however, we took all those reactions happening inside and outside our bodies for granted.

With the emergence of modern science in the middle of the 16th century, scientists started asking more questions about why things work the way they do [9]. By the beginning of the 18th century, the processes of digestion of meat by stomach secretions and starch to sugars by saliva and plant extracts were described, but their mechanism remained unknown [10]. The first enzymes were discovered in the 1830s; they were extracellular, hydrolytic enzymes such as diastase (catalysing breakdown of starch), pepsin (degrading proteins), or invertase (hydrolysing sucrose) [11]. Around the same time, Jöns Jacob Berzelius introduced the concept of catalysts, as chemicals facilitating a reaction without undergoing any change themselves, and hypothesized that enzymes were such catalysts [12]. With more enzymes being discovered, scientists started looking into how the reactions catalysed by enzymes are conducted. It was noted that enzymes are very specific - active on a narrow scope on substrates: pepsin could digest proteins, but not sugars. To explain this phenomenon, a “lock and key” model was proposed by Emil Fisher in 1894, in which enzymes and substrates possess specific, complementary geometric shapes that fit exactly into one another [13]. This model was later modified in 1954 by Daniel Koshland in his “induced fit model”, which describes enzymes as flexible molecules that are reshaped upon binding of a substrate [14]. To explain how enzymatic reactions are conducted, a kinetic model was proposed in 1913 by Leonor Michaelis and Maud Menten [15], in which enzymatic reaction was divided into two stages: reversible binding of substrate to the enzyme, and catalysis of the chemical reaction ending with the release of the product (Figure 1). Their model, in which the rate of enzymatic

reaction is related to the concentration of substrate (Figure 1), is still largely applied to characterise single substrate biochemical reactions.

$$E + S \rightleftharpoons ES \rightarrow E + P$$

$$v = \frac{V_{max}[S]}{K_M + [S]}$$

Figure 1. The Michaelis-Menten model (upper) and equation (lower). Enzyme (E) and substrate (S) are reversibly combined to form the enzyme-substrate complex (ES), which then releases a product (P) and regenerates the original enzyme. The equation relates the rate of product formation (v) to concentration of a substrate. V_{max} represents the maximum rate, happening at saturating substrate concentration. K_M, the Michaelis constant, represents the substrate concentration at which the reaction rate is half of V_{max}.

For a long time, the chemical nature of enzymes was uncertain. Many observed that enzymatic activity is associated with proteins, although some argued that proteins are only carriers, not executors of the activity [16]. In 1926 James B. Sumner crystallized the first enzyme, urease, and confirmed it was a protein [17]. Proteins at that time were recognised as molecules able to flocculate or coagulate under treatment of heat or acid. Gerardus Mulder found in 1838 that nearly all proteins have an empirical chemical formula of $C_{400}H_{620}N_{100}O_{120}P_1S_1$ [18]. Franz Hofmeister and Hermann E. Fischer discovered in 1902 that proteins are polypeptides - they consist of amino acids linked by bonds [19]. Advances in protein purification and crystallization, together with development of X-ray crystallography, allowed for solving their first structures in 1958 [20]. Structures of enzymes followed, confirming the induced fit hypothesis of substrate binding, and opening doors for investigations of how enzymes work on a molecular level. To collect the new protein structures, the Protein Data Bank was founded in 1971 with seven entries, and it grew to over 180000 structures in 2021 [20].

As more and more enzymes were discovered and characterised, the need for structuring enzyme terminology also appeared. In 1961 the Enzyme Commission (EC) classification system was introduced and contained 712 entries in its first edition [21]. The EC numbers consist of four numbers and describe reactions catalysed by an enzyme (Figure 2) Currently there are 7 main classes of enzymes (oxidoreductases, transferases, hydrolases, lyases, isomerases, ligases, translocases) and 6581 reactions with an EC number assigned [22].

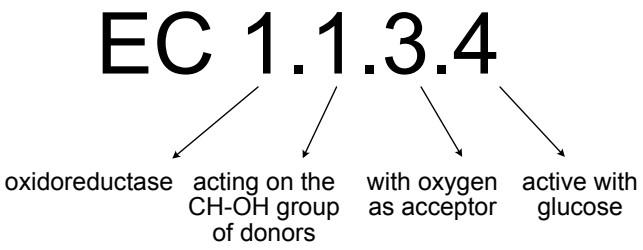

Figure 2. Enzyme Commission classification. Each EC number consists of four numbers describing a progressively more defined classification of an enzyme. Glucose oxidases have an EC number of "1.1.3.4", which means it is an oxidase (EC 1) acting on the CH-OH group of donors (EC 1.1), with oxygen as acceptor (EC 1.1.3), active with glucose as a substrate (EC 1.1.3.4).

In the 1970s commercial application of enzymes began, with glucose isomerase used in production of high fructose corn syrup, and penicillin acylase for production of semi-synthetic antibiotics [6]. However, industrial use of enzymes was largely limited by low quantities they could be obtained in. Similarly, early scientific studies were focused on those enzymes that could be easily obtained in large quantities, for instance by purification from egg whites or blood. During the late 1950s the Armour Hot Dog company purified 1 kg of bovine ribonuclease A and donated it to scientists around the world, which made the enzyme a model for protein studies for the following years [23]. What made proteins much more accessible and allowed for studies of a whole spectrum of new enzymes, was crossing the pathways of enzymology with genomics and molecular biology.

1.2. History of genomics

The existence of discrete inheritable units was first suggested in the 1860s by Gregor Mendel, who conducted experiments by crossbreeding pea plants. It was not until 1952 and the experiments by Martha Chase and Alfred Hershey that DNA was confirmed as a carrier of genetic information [24]. At that time, it was known that DNA consists of four nucleotide bases: adenine, cytosine, guanine, thymine, and their pairing pattern was suggested (A always pairs with T and C always pairs with G). Based on this knowledge, as well as X-ray diffraction data by Rosalind Franklin, James Watson and Francis Crick proposed the double-helix structure model of DNA in 1953 [25]. Five years later a "central dogma of molecular biology" was coined by Francis Crick [26]. It described a flow of genetic information from nucleic acid to nucleic acid or protein and is nowadays commonly referred to as "DNA makes RNA, and RNA makes protein" (Figure 3). By 1967, thanks to work by groups of Marshall Nirenberg, Har Gobind

Khorana and others, the “code for life” was cracked. It was demonstrated that the four bases that build DNA are read in blocks of three to code for a specific amino acid (Figure 3) [27]. With this, the knowledge gap between DNA and proteins was beginning to close.

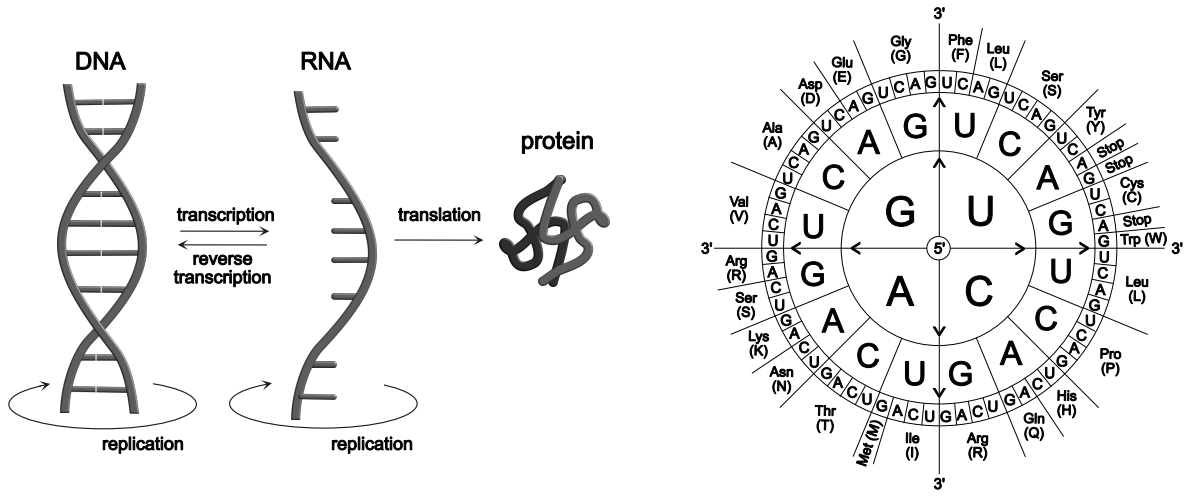

Figure 3. Linking DNA sequence to protein sequence. Left: the “central dogma of molecular biology” describes the flow of genetic information from nucleic acids to nucleic acids or proteins. Right: the genetic code. Codons consist of three RNA bases and encode one of the 20 amino acids or stop signals. The chart is read from the inside outwards, e.g. the codon “AUG” encodes methionine.

In 1972 Paul Berg produced the first “recombinant” DNA by merging a piece of viral DNA into a DNA of the bacterium *Escherichia coli* [28], and so a method of molecular cloning was born. In 1977 with the work of Frederick Sanger, a new milestone appeared - the possibility to determine the nucleotide sequence of DNA, followed by sequencing the first DNA genome - the full genetic information of an organism [29]. Sanger sequencing, although ground-breaking, was time consuming and costly, particularly when applied to whole genomes. At the beginning mainly small and simple viral genomes were published, while the work on more complex genomes was progressing slowly; in 2000, over 20 years after the invention of DNA sequencing, there were nearly 1300 viral genomes published, but only 28 prokaryotic and 6 eukaryotic ones [30]. In 1990 the Human Genome Project was launched to obtain the genetic blueprint of humans. Their first draft of the human genome was published in 2001 [31] and two years later its final version was released. The project was a huge initiative: it took nearly 13 years and cost $2.7 billion. Less than twenty years later, it is now possible to sequence a human genome in one day for $1000 [32]. This drastic decrease in time and money spent was achieved thanks to the “next generation” sequencing methods. These methods started to appear in 2005 and differed in adopted strategies, but all offered short-read, high-throughput, massively parallel platforms [33]. In 2009 the “next generation” methods were complemented

by the “long-read” sequencing strategies that enabled reading ultra-long stretches of DNA, up to 2 million base pairs [34]. When the new sequencing methods became commercially available, the cost of sequencing began to plummet, and the number of genomes sequenced started to increase rapidly (Figure 4).

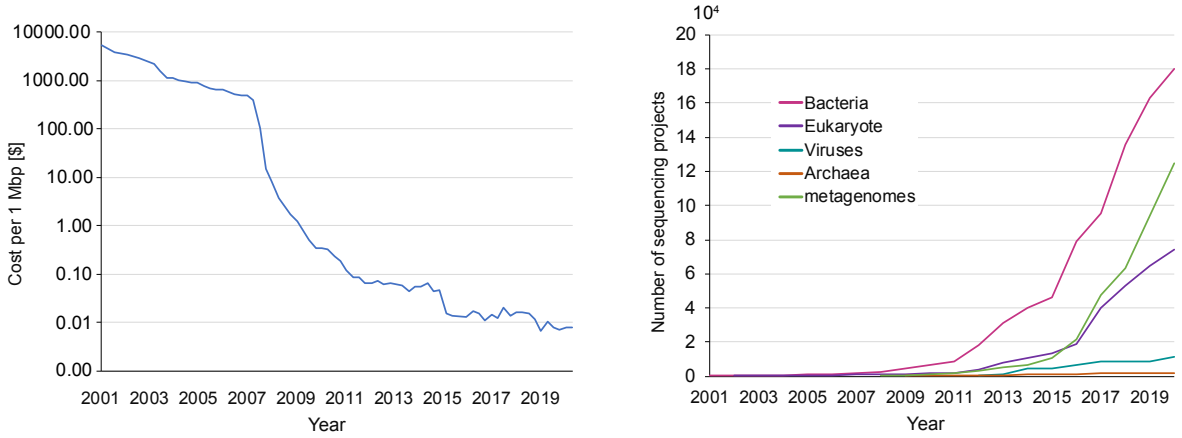

Figure 4. Decrease in the cost of DNA sequencing caused a rapid increase in the number of genomes sequenced. Left: cost of DNA sequencing per one megabase, source: National Human Genome Research Institute [35]. Right: number of sequencing projects deposited to the Genomes Online Database [36].

Suddenly, the scientists were faced with an amount and depth of genomic data never seen before. Although this offered great possibilities, it also came with its challenges, most importantly how to connect sequences of such a huge number of genes with their functions. As it became impossible to elucidate the function of all genes experimentally, the annotation of newly sequenced genes became primarily a computational task. Although not without their issues, the automatic pipelines allowed for processing and annotation of a vast amount of genetic data produced in sequencing projects [37].

1.3. Sequence space of enzymes

All the developments in genetics, molecular techniques, and bioinformatics revolutionised protein science and enzymology (Figure 5). At the end of 1970s it was possible for the first time to read a sequence of a protein, clone it into an easy-to-handle host organism, and produce in large enough amounts to characterise it. Studies on metabolism and biocatalysis accelerated. Recombinant enzymes became very important tools for the development of molecular biology techniques [38]. New enzymes started to find their industrial use: for pharmaceutical synthesis, food production, detergents, personal care products, biofuels, and many others [39]. Discovery of enzymes from extreme environments allowed for exploration

of variants functional in very high or very low temperatures, high salt or acid concentrations [40]. The genomic revolution has provided scientists with an avalanche of protein sequences. Many of these sequences carry unknown functions which might be important for understanding ecosystems around us and prove useful for human applications.

The sequence space of enzymes - all the existing enzyme sequences - is vast. With the possibility of modifying and engineering proteins, the natural sequence space becomes even larger, hiding many possibilities and opportunities. In this thesis I aim to explore various reasons, challenges, outcomes, and approaches of investigating enzyme sequence space. The thesis is based on four manuscripts in the creation of which I was involved, as well as the literature that inspired me in my work. Chapter 2 focuses on describing exploration of the natural sequence space of enzymes, including annotation of enzymes to databases (Paper I), discovery of novel enzymes (Paper II), as well as methods used for large-scale activity profiling of enzymes (Papers I, II, and III). Chapter 3 presents the topic of engineered sequence space exploration, including machine learning-enabled enzyme design (Paper III). Chapter 4 describes methods for enzyme activity profiling, particularly for high-throughput enzyme kinetic measurements (Paper IV). The final chapter presents my outlook on the future of protein sequence space exploration.

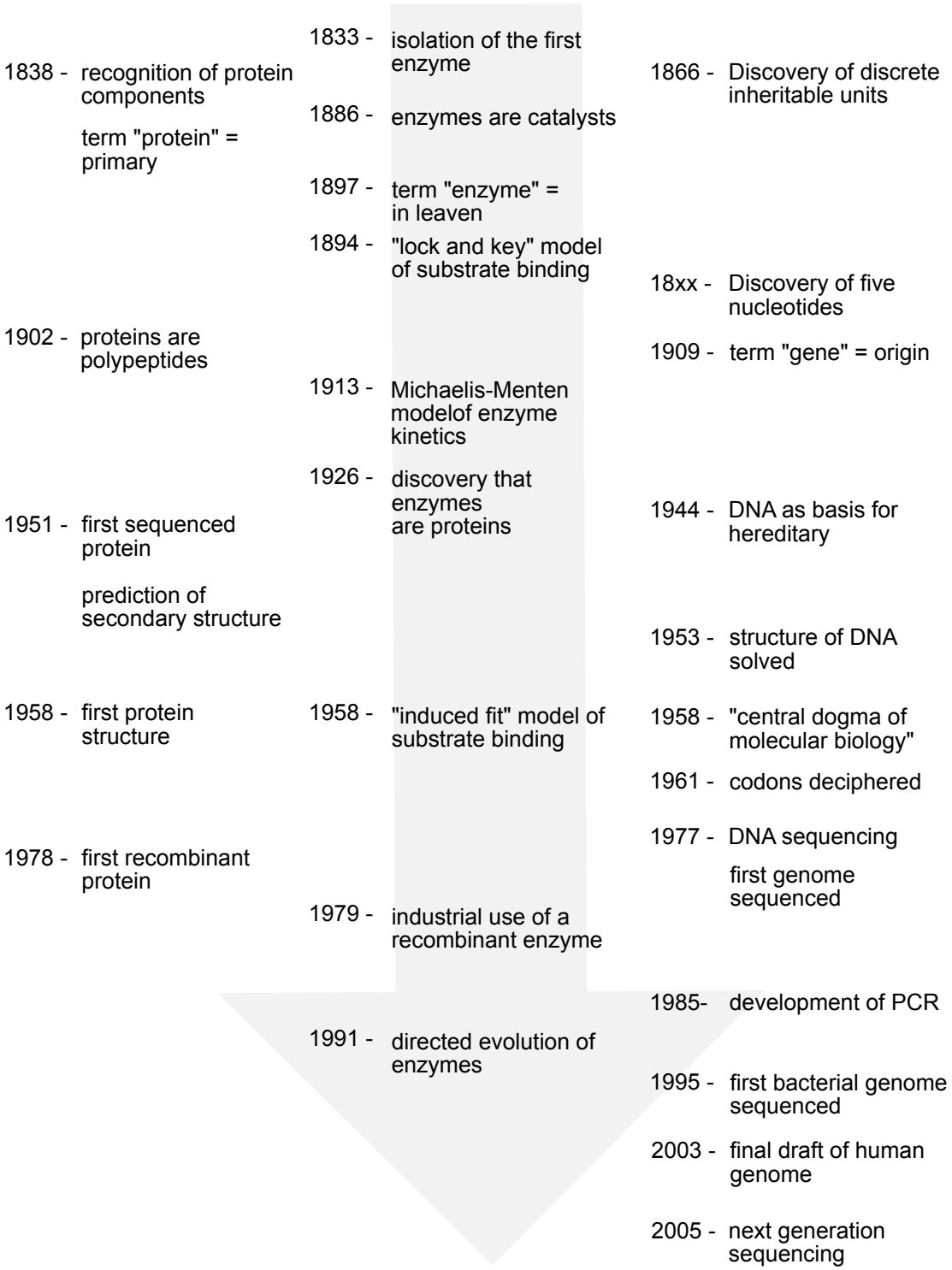

Figure 5. Timeline of major developments in protein research, enzymology, genetics and molecular biology techniques. Inspired by [6].

Chapter 2. Investigating sequence space of natural enzymes

2.1. Underexplored enzyme sequence space

Thousands of different enzymatic functions are present in nature. Some are central to nearly all organisms, while some are extremely specific and present only in a handful. Regardless of their function, they all have one thing in common: they are built of the same building blocks, amino acids. There are only 20 standard amino acids encoded by the genetic code (Figure 3). They are bound together by peptide bonds, and the sequence they appear in in the protein defines their folding into 3D folds, which in turn defines the protein's function. Therefore, structure and function of a protein is encoded in the protein sequence. As more genetic information was collected, it was noted that proteins carrying the same function in two different organisms usually display very high similarity in amino acid sequence and 3D structure. This paradigm became a basis for homology-based computational prediction of protein function, where newly sequenced genes have their function assigned based on their sequence similarity to characterised proteins. In enzymes, a few amino acid residues are essential to carry out a catalytic function, such as the ones involved in substrate or cofactor binding. Conservation of these proved to be more crucial in carrying out a function than overall similarity of the whole sequence [41]. Thus, searching for "motifs" - characteristic signatures of a fold or domain associated with a function - became another method for function prediction. More refined methods of functional annotation rely on structure similarity, genomic context, or phylogeny [42].

Computational methods for functional annotation, although constantly improving, often provide only broad clues for gene functionality, rather than specific answers. Prediction of completely novel functions is particularly challenging [43]. Additionally, incorrect automatic predictions are very common and tend to percolate, leading to error accumulation in the databases [44]. Experimental approaches provide the highest level of information for elucidation of enzymatic functions and create a base for reliable computational annotations. However, the classical biochemical methods of enzyme characterisation cannot keep up with the amount of genomic data produced. As such, protein studies tend to be heavily biased towards characterizing proteins from eukaryotes (Figure 6), model organisms, or disease-causing bacteria [45]. A

recent study revealed that enzymes from eukaryotes, the least diverse part of the tree of life, comprise 55% of all characterised enzymes, and five mammals contribute to 15% of enzyme database entries [46]. A large number of bacterial and archaeal phyla was shown to be completely underexplored. These data clearly show that a big part of the natural sequence space of enzymes remains underexplored.

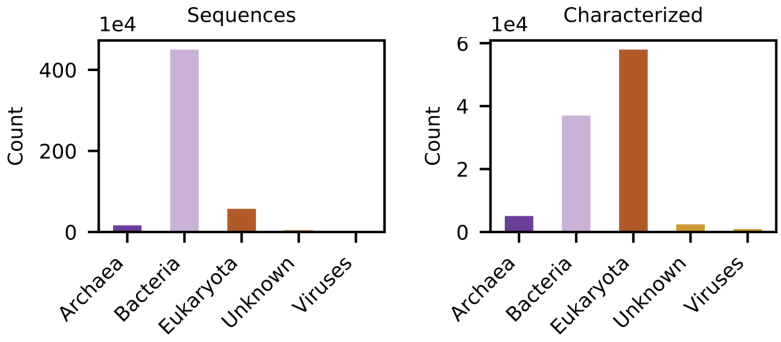

Figure 6. Experimental bias of protein investigation. Although there are more bacterial than eukaryotic enzymes deposited to the BRENDA enzyme database (left), the number of experimentally investigated bacterial enzymes is much smaller than eukaryotic ones (right). [47]

2.2. Large-scale activity profiling of enzymes

2.2.1. The subject of large-scale activity profiling

The issues with traditional experimental and computational methods for function annotation have led researchers to look for alternative ways to explore enzymes. Large-scale “genomic enzymology” approaches aim to shift the focus of enzyme exploration from a host organism to sequence space, enabling sampling through a wider swath of functional space. In such approaches, computational screening of sequences is followed by a high-throughput experimental characterisation of selected candidates with a large scope of potential substrates. Most commonly the subject studied with the high-throughput sequence space approach is an enzyme family, a group of evolutionary related proteins, often displaying sequence, structure, and function similarity. One of the earliest large-scale profiling experiments were carried out to unravel functions of families with domains of unknown functions [48,49]. Other subjects were experimentally underexplored families [50–55] or non-homologous isofunctional enzymes [56]. In contrast, the subjects of studies in Paper I and Paper II are enzyme classes, rather than enzyme families. Enzyme class, described by an EC number (Figure 2), groups enzymes based on the reaction they catalyse, not homology. This means that two enzymes which catalyse the same reaction but are members of two evolutionary unrelated protein families would be assigned an identical EC number. Applying large-scale activity profiling to enzyme classes allows for exploration of many different sequences catalysing similar reactions and experimental validation of their existing annotations. Additionally, when applied to several enzyme classes and a wide range of different substrates, the experimental platform is also well-suited for discovery of promiscuous activities - side reactions that are distinct from the enzyme's main activity.

2.2.2. Workflow of large-scale activity profiling

Regardless of the subject of study, performing a large-scale activity profiling follows a three-step procedure: selection of candidate sequences, protein production, and activity screening (Figure 7).

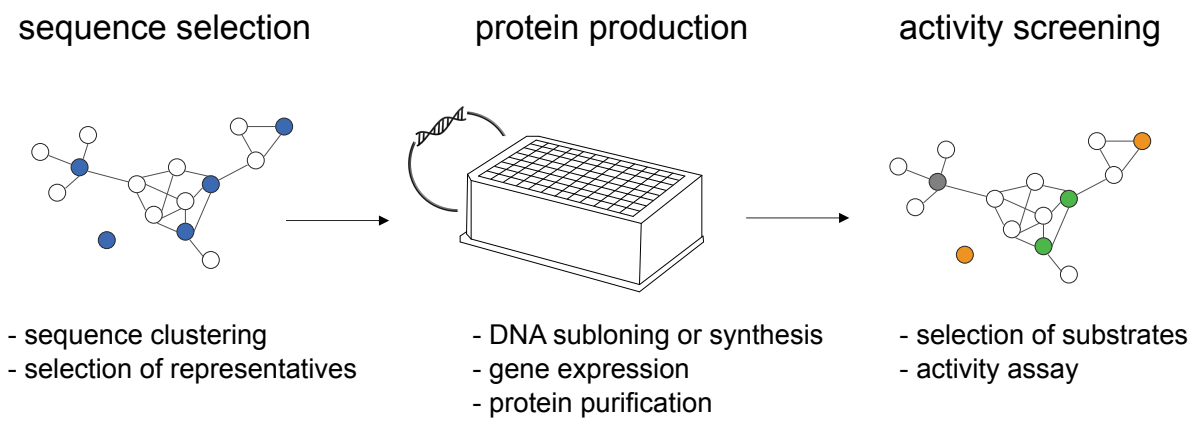

Figure 7. The typical workflow of large-scale activity profiling of enzymes.

The first part of sequence selection often involves creating a sequence similarity network (SSN), where subfamilies of similar sequences are clustered together. In most studies the candidate sequences are sampled throughout the whole sequence space to represent its diversity as best as possible. Sometimes a particular weight is put on characterisation of sequences from underexplored or extremophilic hosts, very distant family members or hypothetical enzymes [53,54,56]. In Papers I and II sequences were iteratively selected so that each newly chosen sequence maximally increases the mutual information explained within each cluster.

Once the candidate sequences are selected, their production comes next. Usually, hundreds of sequences are chosen in the first step, so the enzyme production platform has to be adapted for a higher throughput. In the early large-scale activity profiling studies selected genes were subcloned from the host genome, which involved a lot of hands-on time and occasionally limited the scope of selection in step 1, as it was dependent on availability of genomic DNA. In the most recent studies, including Papers I and II, gene synthesis is preferred over subcloning, due to its convenience and possibility of codon optimisation. *E. coli* is usually chosen as an expression host of choice because of its ease of growth in high-throughput conditions. Expression of genes is carried in 1 - 2 ml culture volumes in multiwell plates, and affinity purification is conducted in a 96-well format, if necessary. After production, the presence of enzymes is assessed by SDS-PAGE or its automated, capillary-based alternative LabChip GX. Protein expression and/or solubility is often a limiting factor in large-scale activity profiling, with recovery ranging 15 - 75%, which can lower greatly the number of proteins that can be tested in the downstream analysis. This could be minimised by solubility prediction at the sequence selection step [53] or codon optimisation for the expression host.

Functional screening is the final, and most crucial step. Like enzyme production, it should be easily adaptable for a high-throughput setup. Oftentimes selection of the subject of study relies on the availability of a high-throughput screening method for a given enzyme family or class. The most common activity screening methods are based on measuring change in absorbance or fluorescence of a substrate, product, or a coupled probe in a multiwell plate. The successfully purified enzymes are assayed with a range of pre-selected substrates and their activity recorded by endpoint or continuous measurements. The initial results of this large-scale activity profiling are usually just a stepping-stone to further, more detailed studies. Depending on the obtained results, investigations may continue by testing a broader range of substrate and activities, analysing sequence-function relationship within the enzyme sequence space, or obtaining detailed kinetic, biophysical, or structural information for selected candidates.

2.2.3. Exploration of EC 1.1.3.x in the “all-vs-all” experiment

The first application of a large-scale enzyme profiling for characterisation of an entire protein family was performed to assign a function to proteins containing a domain of unknown function [48]. The study discovered 14 new β-keto acid cleavage activities, unravelled their metabolic functions and predicted key residues responsible for specificities. Since its publication, each year new studies using similar approaches and tools are being released, expanding knowledge of sequence-structure-function relationships of the investigated enzymes. In Papers I and II a high-throughput experimental platform was used to screen a sequence space of selected enzyme classes. We chose oxidases acting on the CH-OH group of donors with oxygen as an acceptor (EC 1.1.3.x) as a subject of the studies. Nearly 15000 sequences are currently annotated as 1.1.3.x in UniProt (https://www.uniprot.org/), but only a fraction of them have been experimentally characterised. Representative of EC 1.1.3.x take part in a wide range of biological processes, such as ascorbic acid production [57], photorespiration [58], production of osmoprotectants [59], synthesis of antibiotics [60,61] and phytotoxins [62]. They are also of interest to the medical and food industries as biosensor candidates for sensing specific marker molecules, like glucose, lactate, ethanol, cholesterol, galactose, and choline [63,64]. Additionally, some EC 1.1.3.x oxidases display a big potential in organic synthesis, being used in production of drugs, antioxidants, flavour and fragrance compounds [65,66]. The majority of 1.1.3.x oxidases produce hydrogen peroxide as one of the products, which

can be easily detected in a fluorescence-based assay (Figure 8). The ease of assaying, adaptability for a high-throughput setup, as well as their biological and industrial relevance makes the enzyme class a perfect candidate for large-scale activity screening.

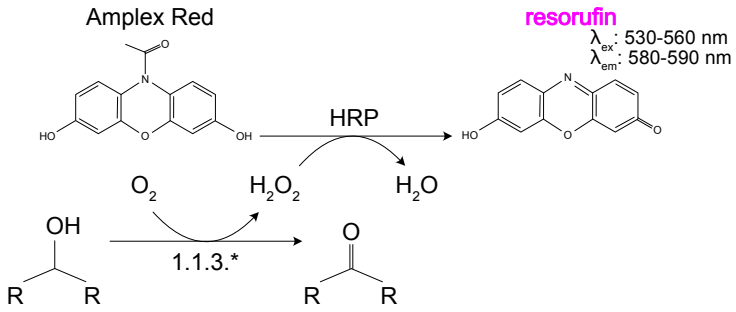

Figure 8. General reaction schemes of 1.1.3.x oxidase (bottom) and the Amplex Red activity assay (top) used for hydrogen peroxide detection, and applied for assaying enzymes in Papers I, II and IV.

In order to get an overview of the activity profile of the oxidases selected for my work, we first conducted an "all-vs-all" experiment. 96 candidate enzymes annotated to EC 1.1.3.x were chosen to be screened with 23 representative substrates of EC 1.1.3.x for the oxidation activity (Paper II, Tables S1 and S2). Typically for a large-scale screening platform, we experienced a drop in numbers of produced proteins, with 58% recovery, which left us with only 54 enzymes to screen. Surprisingly, the vast majority of the purified enzymes displayed no activity, even with their predicted substrates. Part of this inactivity could be explained by improper folding of the produced enzymes, a limited number of substrates tested, or assay conditions. However, in a follow-up study we confirmed that the main reason behind the result was the issue of misannotation in enzyme databases (Paper I). Additionally, the "all-vs-all" screen revealed two enzymes active with substrates of other enzyme classes than the predicted ones. The discovery and characterisation of the two novel oxidases were described in Paper II.

2.3. Exploration of functional annotations in enzyme databases.

2.3.1. Protein databases

As proteins began to reveal their vast repertoire of functions, new databases were being created to accommodate the collected data. Primary databases, such as GenBank [30] and PDB [67], started collecting direct results of sequencing or structural projects, and serve primarily archival purposes. Secondary databases, otherwise known as knowledge databases, began collecting experimental results, literature research, or interpretation of data from other databases. One of the largest secondary protein databases is UniProt, consisting of Swiss-Prot and TrEMBL resources [68]. Swiss-Prot was established in 1986 to contain manually curated entries and information on protein function, domain structure, modifications, variants, literature, and links to other secondary databases, among others. Manual curation is a labour-intensive undertaking, and ten years after Swiss-Prot its non-curated supplement, TrEMBL, was created to accommodate computationally annotated protein entries [69]. BRENDA, created in 1987, is one of the largest enzyme-focused databases, collecting functional and molecular information about enzymes from primary literature [70]. A whole range of specialised enzyme-related databases exist, with focus on pathways and metabolism, such as KEGG [71] or MetaCyc [72], as well as those collecting information on specific enzymatic functions, such as CAZy (carbohydrate-active enzymes) [73], or PeroxiBase (peroxidases) [74].

The majority of protein databases contain both high-quality functional annotations based on the literature or manual curation, as well as computational annotations, often transferred directly from primary databases or UniProt. Functional annotation is not a straightforward task, and errors often creep in. Firstly, issues may appear at the genome sequencing and assembly step: errors in gene sequencing, border establishing, or assembly inevitably result in incorrect protein sequences deposited to databases [75]. Secondly, there are issues with the functional prediction of the protein function itself [76]. Oftentimes the presence of motifs and domains is used as a substitute for a functional assignment, where a sequence with a similar set of domains to known protein is automatically assigned with this protein's function. It is common to see specific assignments with no, or very remote, similarity to proteins of known function. Additionally, terms like "predicted" or "possible" function are used that do not carry much information about the actual function. Single errors of functional annotation can have disproportionate consequences. Chains or "percolations" of misannotations appear, as the

annotations are often copied from similar protein sequences in each database, not necessarily from experimentally characterised or curated sequences [44].

It is hard to estimate the exact extent of functional misannotation in protein databases. One of the best attempts to do this was published in 2009 by Patricia Babbitt's group [77]. In the study by Schnoes et al., functional annotation of 37 well-studied enzyme families was evaluated in one primary database (GenBank) and three secondary databases (Swiss-Prot, TrEMBL, KEGG). The manually curated database, Swiss-Prot, had by far the lowest annotation error - close to 0%. The other three databases displayed a similar, and surprisingly high error: 5-80%, depending on the enzyme family. The most common annotation error was "overprediction", where sequences were members of a superfamily, but there was not enough evidence that they catalyse a specific enzymatic reaction. It is important to point out that this study, as well as other investigating the issue of functional misannotation, were published before the "sequencing boom" (Figure 4), when the size of genome and protein databases was much smaller [77–79]. Awareness of the issue has been raised, and new annotation methods have developed, so has anything changed since then?

2.3.2. Misannotation of enzyme classes to enzyme database (Paper I)

In Paper I we performed a computational and experimental analysis of the sequence space of enzymes annotated as S-2-hydroxyacid oxidases (EC 1.1.3.15) in the BRENDA database [47]. This enzyme class is the largest one among the 1.1.3.x oxidases; at the time we started the work 1058 unique sequences were annotated to it. Representatives of the class oxidize S-2-hydroxyacids like glycolate or lactate to 2-oxoacids. Characterised members of the EC 1.1.3.15 operate on a broad substrate range *in vitro*, although the physiological substrate for plant and mammalian homologues is mainly glycolate or long chain hydroxyacids [58,80,81], while lactate is the main physiological substrate of bacterial homologues [82,83]. Members of EC 1.1.3.15 are of high biological importance, with plant glycolate oxidase being crucial for photorespiration, mammalian hydroxyacid oxidases taking part in glycine synthesis and fatty acid oxidation, and bacterial lactate oxidases metabolising L-lactate as an energy source [58]. The latter are of particular medical and industrial interest, being used for lactate biosensor development in clinical care, sport medicine, and food processing [84].

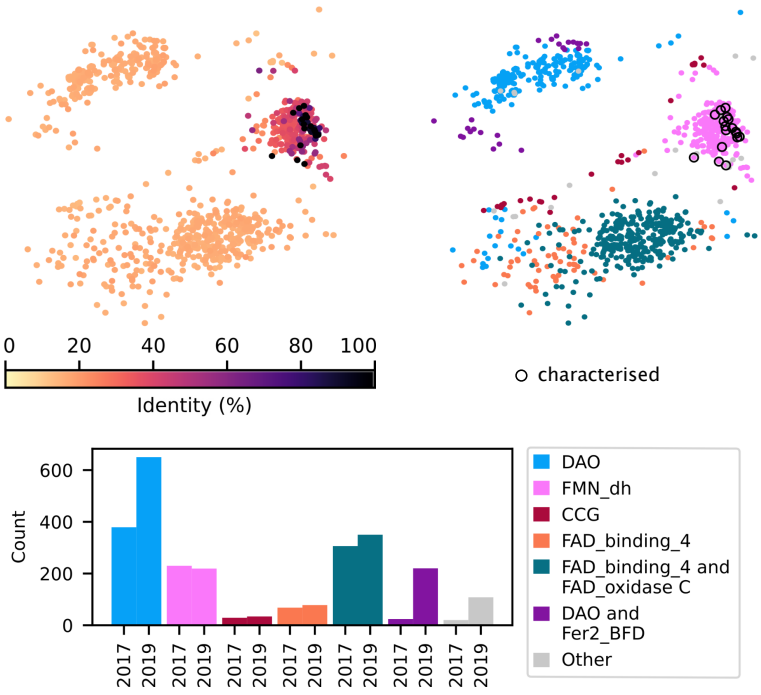

Figure 9. Sequence space of enzymes annotated to EC 1.1.3.15, where proximity between two points indicates sequence similarity. Top left: percentage of sequence identity to the closest experimentally tested or curated S-2-hydroxyacid oxidase. Top right: Pfam domain architecture. Bottom: comparison of Pfam domains of sequences annotated to EC 1.1.3.15 in BRENDA version 2017.1 and 2019.2. Adapted from [47].

The initial bioinformatic analysis of the sequences annotated to EC 1.1.3.15 showed that nearly 80% of them share less than 25% sequence identity with characterised representatives of the family and have a different predicted domain architecture (Figure 9). This result by itself is a very strong indication of misannotation of those sequences to the enzyme class. To further investigate the large diversity of the sequences, we performed a high-throughput experimental validation of their predicted activity. 122 candidate sequences were selected throughout the EC 1.1.3.15 sequence space for the large-scale activity profiling. 65 proteins were successfully produced and assayed with six different 2-hydroxy acids. In the cluster of sequences homologous to the known S-2-hydroxyacid oxidases, the majority of enzymes displayed a typical activity profile for the oxidases (Figure 10). The inactive sequences within the cluster were primarily missing one or more of the amino acids crucial for catalysis. Additionally, we found a few sequences in the homologous cluster that were most likely members of the same superfamily as S-2-hydroxyacid oxidases but catalysing different reactions. Through experimental profiling, we also confirmed that the sequences with little or no similarity to the characterised S-2-hydroxyacid oxidases are indeed misannotated: are either inactive or do not display a broad substrate scope characteristic for EC 1.1.3.15 (Figure 10).

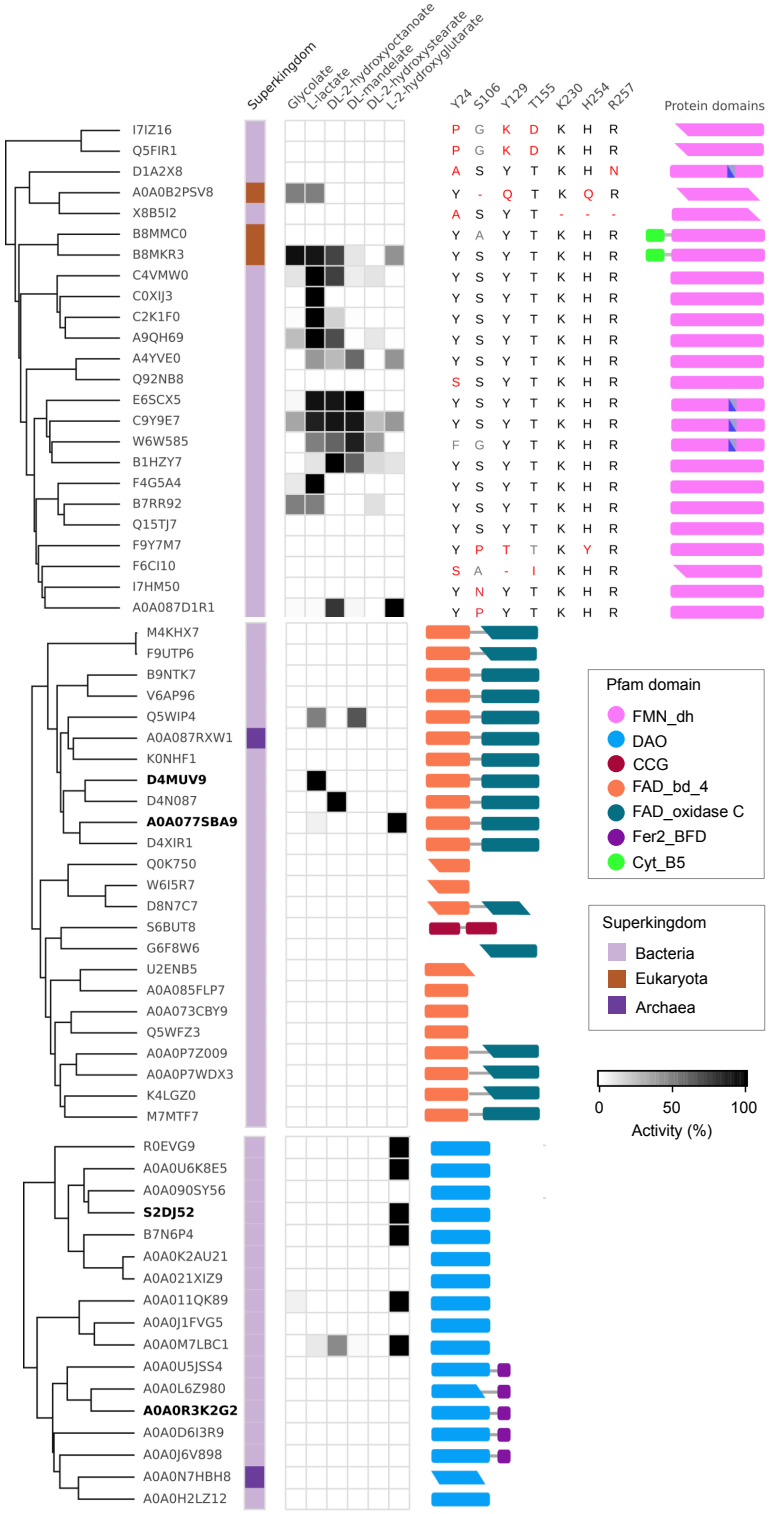

Figure 10. Experimental profiling of sequences annotated to EC 1.1.3.15. Dendrogram indicates protein relatedness. Recorded activities are marked with squares, for proteins active with more than one substrate, the substrate preference is shaded with the highest activity for each enzyme scaled to 100%. Listed amino acids correspond to conserved residues in a glycolate oxidase from *Spinacia oleracea.* The cartoons represent predicted domains and motif composition of the sequences, based on Pfam search. Domains lacking full Pfam alignment are represented with a sharp edge. Proteins with alternative activities chosen for kinetic characterisation are marked in bold. Adapted from [47].

We were not able to establish activities for all the misannotated sequences, however, we did confirm the presence of enzymes with four alternative activities among them: L-2-hydroxyglutarate dehydrogenase, D-2-hydroxyglutarate dehydrogenase, D-lactate dehydrogenase, and glycerol-3-phosphate dehydrogenase. Like in previous studies investigating misannotation in databases, we confirmed that the misannotation in this enzyme class accumulated over time, from at least 78% in 2017 to 87% in 2019 (Figure 9).

To find out whether the large annotation error is only a problem of this particular enzyme class, we analysed annotations into all enzyme classes in BRENDA, based on sequence and domain similarity to characterised and curated sequences (Figure 11). Strikingly, almost one fifth of sequences share less than 25% pairwise sequence identity with the closest characterised or curated enzyme of the enzyme class they were annotated to. Additionally, 18% of all sequences, mainly the low-identity ones, are not predicted to have the same Pfam domains as the experimentally characterized enzymes. Based on the results obtained for 1.1.3.15, these sequences are very likely to be misannotated. Although this analysis was performed on the BRENDA database, we expect that similar levels of misannotations are present in other databases, since BRENDA, like many other secondary protein databases, largely relies on annotations from UniProt.

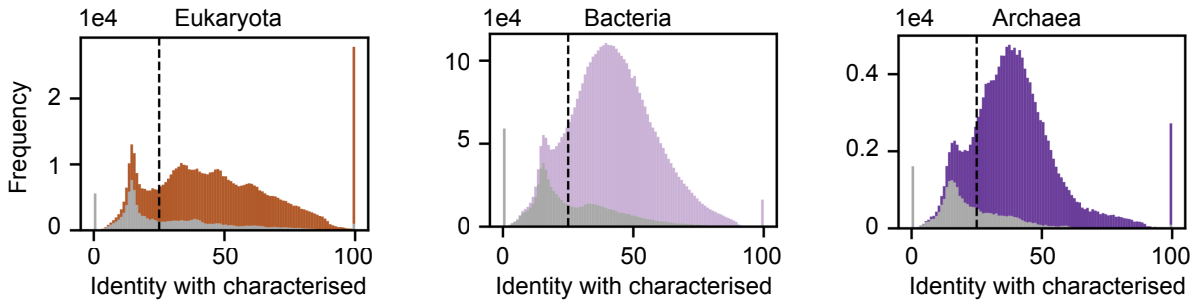

Figure 11. Misannotation to enzyme classes in BRENDA DB. Histograms show the distribution of sequence identities between sequence cluster representatives (after clustering at 90% identity) and their closest characterised/curated enzyme for Eukaryote, Bacteria, and Archaea. Proteins which do not have the same Pfam domains as characterised/curated enzymes are coloured in grey. [47]

Just like other studies investigating annotation errors in protein databases, Paper I showed that misannotation continues to be a big issue. The main difference between the earlier findings and ours is the reason for the error in annotation. In the work by Schnoes et al. [77], which was based on entries to public databases in 2005, only 3% of all sequences were considered misannotated due to the lack of sequence similarity and domain architecture to

the characterised representatives. Although using a slightly different methodology for error estimation, we show in our study that 15 years later this number is much higher now - close to 20%. One can think that these kinds of annotation errors would be very easily fixable even with the most basic homology-based annotation methods. The problem, however, lies in the way genome annotation pipelines work [44]. New entries to genomic databases are usually annotated based on a “majority rule” - similarity to already existing entries, with little concern as to what the previous annotations were based on. This can result in a sequence being annotated not based on similarity to the closest characterised sequences, but on similarity to the largest number of already annotated sequences – whether the annotations were carried correctly or not [85,86]. Genome annotation pipelines largely follow a policy of providing maximum coverage of annotation, rather than as good annotation as possible based on the current experimental and curated data. Instead, identity to the nearest characterised sequence combined with prediction of domain architecture should be vital checkpoints in functional annotations of genomes. Many secondary protein databases, such as UniProt, InterPro, or KEGG already use more refined automated systems for functional annotations, some of which contain curated sequences as templates. However, when asked about their annotation policies, the UniProt representatives admitted that application of such systems is often limited and does not cover all the sequences, which might also be true for other databases.

Similarly to previous misannotation studies, in Paper I we also found the types of annotation errors which cannot be easily corrected with the simple homology-based methods: sequences belonging to the same superfamily but not the same enzyme class, or those without functionally important residues. The more sophisticated annotation methods could be of help here, although they too have their limits. Additionally, their incorrect use might result in overprediction of function, for instance when a presence of a domain or a motif is made equal to the presence of a certain function [86]. When no clear evidence for a functional assignment exists, a sequence should be annotated only with the level of information one can confidently assign to it, for instance as a member of a superfamily [77]. This approach avoids spreading misinformation and provides a good starting point for guiding experimental annotation efforts. However, with such an approach many sequences are annotated only with their general characteristics, and the number of sequences for which reaction specificity is annotated is greatly lowered. Although this may result in less densely annotated genomes, their overall quality will be of much higher standards.

2.3.3. Registration of experimental data

Our article, like many others, proved the value of experimental validation of annotations. What is equally important, but often overlooked, is correct registration of experiments. The classical biochemical characterisation studies cannot keep up with the amount of genomic data produced. It is therefore of prime importance that the ones that are being performed, could become a basis for computational annotations. This might seem like an obvious statement but is still an issue. In Paper I we described four proteins which were misannotated as S-2-hydroxyacid oxidases, but for which we found alternative activities. After a literature search, we found articles describing close homologues of all the four proteins, which carried the same activities as we found [87–90]. Only one of the articles proposed an annotation transfer [87] which resulted in a recent re-annotation of the protein in UniProt. The remaining enzymes are still not recorded in protein databases as being experimentally tested, and as such do not serve as a reliable base for function transfer. In those three articles the main issue was the lack of a clear link between the characterised enzyme and its sequences. It is not uncommon when describing a new activity that only the gene name or abbreviation is stated, which is not enough to identify the unique enzyme sequence. As a response to this issue, the journal *Biochemistry* called on authors to include unique accession identifiers, such as UniProt or NCBI IDs, of all the proteins characterised in the manuscript [91]. This requirement facilitates capture of experimental data by electronic search engines, for example of UniProt, and should certainly be adopted by other journals. Submission of new experimental data directly to databases is also possible: secondary protein databases, such as UniProt or BRENDA, welcome users' corrections, however, it is uncertain to what extent those options are actively used by the community and result in correction of annotations. With the rise of large-scale activity profiling approaches, there is also a need for a discussion about registering such results in databases. High-throughput screening platforms enable obtaining a large amount of evidence about activity or substrate scopes of tested enzymes. Although not as precise and thorough as the classical characterisation methods, they could nevertheless provide good evidence for computational annotations. The articles which apply high-throughput platforms do excellent work investigating underexplored enzymatic sequence spaces, yet their results are mostly not recorded in protein databases. A structured way of utilising the large-scale activity profiling data should be organised, to make the most of the results. Interestingly, one of the works describing novel carbohydrate-active enzymes has part of the results registered in the CAZy database, but not UniProt, the most common source used by protein feature databases [54]. This indicates the need for more careful cross-reference between databases.

Correct annotation of genes is crucial for exploration of novelty and understanding fundamentals of biological functions. The fields of systems biology [92], metabolic and enzyme engineering [93,94] also heavily rely on accurate functional annotations. One example of the dire consequences of misannotation in protein databases comes directly from our laboratory. A project aiming at testing various thermophilic lactate oxidases for sensor development was conceived. Several sequences with high catalytic temperature optima were chosen for screening from a publicly available dataset [95]. The proteins purified well, but to our surprise none had the expected lactate oxidase activity. After some time of confusion, we realised that the proteins were misannotated, and their sequences had no similarity to the known lactate oxidases. Admittedly, the non-homologous genes should have been filtered out before ordering the sequences, but at this time we did not think about doubting the existing annotations. Time and money were spent, lessons were learned. Similar mistakes might have been made in the laboratories worldwide, although it is hard to precisely estimate to what extent misannotation affects research projects, as mistakes are rarely reported, and some issues are never detected.

Protein databases play a crucial role in collecting information about sequence, function, structure, interactions, among others. They form a basis for ever-evolving annotation methods, as well as provide a starting point for many experimental research. Currently in many databases, right next to high quality information, misinformation spreads. Not all the users might be aware of this issue or are able to judge the quality of information. It is therefore of prime importance that both experimental, and computational data are registered in such databases to the highest possible standards.

2.4. Discovery of novel enzymes

2.4.1. Exploration of the “catalytic dark matter”

One of the reasons why it is difficult to functionally annotate proteins, is because we simply have not discovered all the types and flavours they come in. There is a vast enzymatic potential in nature, but for many activities we still do not know their sequence-function relationships. In all newly sequenced genomes, there is a portion of genes for which no function is annotated, the so-called “genes of unknown functions”. Usually, their sequence is too dissimilar to any other known protein to be classified. Genes of unknown functions make up a sizable portion of even well-studied organisms: 20-30% of human, yeast and bacterial genomes [43,96,97], out of which a third were estimated to be enzymes [43]. A special case of unknown enzymes are orphan enzymes - experimentally characterised enzyme activities, for which we lack sequence information. Around 20% of current enzyme classes do not have corresponding genes linked to them and are considered orphan [98]. These numbers show that there is still a lot to unravel about the mechanisms ruling human bodies and other fellow organisms. The “catalytic dark matter” of plants and microorganisms is of particular interest as a source for enzyme discovery, with many esoteric activities and useful biocatalysis lurking in the less studied genomes [43,99].

So how are novel enzymes found? There are two general ways to approach the task: using hypothesis-driven or untargeted methods [43]. Hypothesis-driven approaches start with screening for a known or theorised activity. Such approaches were commonly used in the beginning of enzyme studies, when crude cell lysate was assayed for an activity of interest and multiple steps of purification led to obtaining fractions enriched for the new activity. Pure preparation can later be identified by protein sequencing or mass spectrometry. This approach, although tedious and time-consuming, is still being used nowadays, and has recently led to the identification of such proteins as lignin-modifying hemocyanin from a termite or a plant glucuronokinase [100,101]. A high-throughput variant of this classical biochemical approach is metagenomic screening for novel functions. Instead of cell lysate, genetic libraries obtained from environmental samples are screened for a desired function. It is particularly popular when screening for industrially relevant biocatalysts and resulted in the discovery of hundreds of novel microbial cellulases, chitinases, oxidases, lipases, among others [102,103]. The most commonly applied hypothesis-driven approaches start with a gene of interest that is subcloned, expressed, and characterised. Bioinformatic tools based on homology or motif

presence are often used to generate functional hypotheses and narrow down what activities and substrates to screen for. High-throughput activity profiling, described in Chapter 2.1 of this thesis, is a variant of this method where the “exploratory net” is cast much wider, in terms of numbers and types of sequences, as well as activities being tested. Hypothesis-driven approaches are still the most commonly used methods for enzyme discovery, although they do come with a caveat of: “you get what you screen for”, where unexpected enzymatic activities might be missed. This problem could be partially solved by untargeted approaches, which begin with little or no prior information about potential substrates or products of the enzyme of interest. In the early days of gene function discovery, many activities were confirmed by studying deletion mutants of model organisms: if a clear phenotype of a mutant was observed, this could be later linked to a gene function. Enzyme discovery using metabolomics is also another method which applies an untargeted approach. Enzymes of interest are purified and incubated with an enriched molecular extract, and their activities are revealed by analysing consumption or production of metabolites using mass-spectrometry. Similar techniques can be applied to study cells deficient in or overexpressing a gene of unknown function. Metabolomics-based approaches can also be scaled-up and has recently led to identification of 241 potential new enzymes in *E. coli*, of which 12 were experimentally validated [104].

Although experimental validation is necessary to confirm a novel function, computational approaches play an equally essential part in driving the discovery of enzymes and oftentimes provide a starting point for experimental research. Comparative genomics methods rely on inferring the function of an unknown gene through its association with known genes and extracting information from biological databases about the genomic context, coexpression, or protein-protein interactions [105]. *In-silico* docking methods can provide clues about enzyme’s substrates and ligands [106]. Genome-scale metabolic networks allow for evaluation of “missing” enzymes by identification of dead-end or disconnected metabolites [107]. New computational tools, platforms and databases aiming at guiding experimental attempts are being created, such as the Enzyme Function Initiative Tools for creating sequence similarity networks and genome neighbourhood maps [108], STRING database collecting functional association networks [109], or PaperBLAST search engine looking for scientific articles about homologous proteins [110]. Initiatives like CAFA (Critical Assessment of protein Function Annotation algorithms) aim to bring the scientific community together in order to assess the quality of current available annotation methods [42].

2.4.2. Discovery of two novel oxidases (Paper II)

Our high-throughput activity profiling of EC 1.1.3.x, described as “all-vs-all” experiment in Chapter 2.2.3 and Paper II [111], is an example that combines the hypothesis-driven and untargeted approaches. In our setup, a number of non-homologous sequences were screened with a range of very different, structurally diverse substrates. As such, this setup presents more opportunities of finding unexpected or side activities than standard activity profiling approaches. Indeed, screening 96 putative oxidases with 23 diverse substrates resulted in discovery of two novel enzymes: an orphan enzyme N-acetylhexosamine oxidase from *Ralstonia solanacearum*, and a novel long chain alcohol oxidase from an uncultured marine euryarchaeote, an example of a non-homologous isofunctional enzyme.

The majority of orphan activities were identified before the sequencing methods were readily available [112]. The enzymes were usually purified from native hosts and thoroughly characterised, including such information as the substrate scope, reaction kinetics, isoelectric point or molecular weight, but no amino acid sequence information was linked to them. One example of such orphan enzyme is N-acetylhexosamine oxidase (HexNAcO), discovered in 1989 in a Pseudomonas-like bacterium [113]. The enzyme was confirmed to be a flavoprotein oxidase active with a range of monosaccharides and displayed the highest activity with N-acetylated hexosamines: N-acetylglucosamine (GlcNAc) and N-acetylgalactosamine (GalNAc). Information about the enzyme’s substrates, reaction, and catalytic optima was included, but no sequence investigation was carried out in the original paper.

In our “all-vs-all” screen we found a protein displaying activity with GlcNAc. We carried out characterisation of the protein and found that its substrate scope is very similar to the one published for the original HexNAcO (Table 1). In this manner, we confirmed that the enzyme from *R. solanacearum* is indeed a HexNAcO, and we found the sequence for this orphan activity. Once available, analysis of its amino acid sequence revealed that the enzyme is homologous to characterised fungal FAD-binding saccharide oxidases, as well as bacterial oxidases involved in antibiotic production. All these proteins are berberine bridge enzyme (BBE) -like enzymes, which display a vanillyl-alcohol oxidase (VAO) fold and contain an atypical bi-covalent anchoring of the FAD cofactor [114] (Figure 12). What seems to distinguish the novel HexNAcO from the other VAO enzymes, is an elongated stretch of amino

acids by the edge of the substrate binding cavity (Figure 12). A similar pattern can also be found in another saccharide oxidase operating primarily on monosaccharides: hexose oxidase from *Chondrus crispus [115]*. It is possible that such an extension acts as a lid of the substrate binding domain, allowing in primarily smaller substrates. Currently ongoing structural investigation of the HexNAcO revealed that even in a high-resolution structure (1.5 Å), this region did not generate an interpretable X-ray diffraction pattern, which might indicate a flexible or disordered nature of the loop.

Table 1. Comparison of substrate specificity of the N-acetyl-D-hexosamine oxidases characterised in the paper first describing the activity (Horiuchi, 1989) and in our paper (A3RXB7). [111]

Substrate	K_M [mM]		Specific activity [μmol min^{-1} mg^{-1}]	
	Horiuchi, 1989[113]	A3RXB7	Horiuchi, 1989[113]	A3RXB7
GlcNAc	0.24	0.26	71	6.08
GalNAc	0.1	0.32	70	4.78
ManNAc	40	182	12	2.39
chitobiose	18	19	45	0.9
D-glucosamine	40	4.5	34	0.38
D-galactosamine	10	1.4	35	0.67
D-mannosamine	--	65	--	0.01
D-glucose	290	216	3.8	0.17
D-galactose	170	102	3.3	0.21
D-mannose	59	118	1.2	0.15

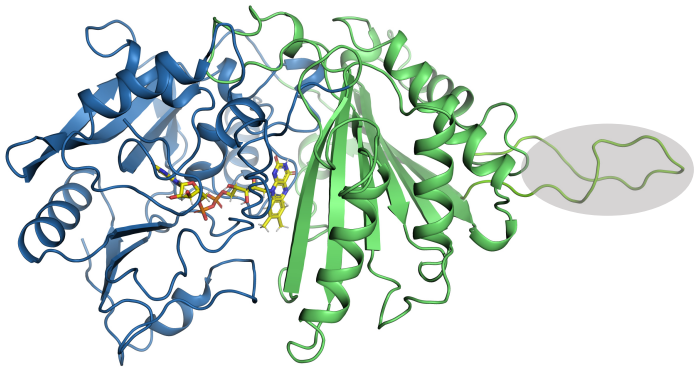

Figure 12. Homology model of HexNAcO. FAD-binding domain is coloured in blue, substrate binding domain is coloured in green, highlighted in grey is the elongated loop of the substrate binding domain. Model obtained using RaptorX server, using 2Y08 chain A structure (tirandamycin oxidase TamL) as a template. [111]

Non-homologous isofunctional enzymes (NISE) are evolutionarily unrelated proteins that catalyse the same biochemical reactions [116]. As they share no sequence, and often no structure similarity, it is impossible to infer a function of one of them from the sequence similarity to the other. In our “all-vs-all” screen we found a protein active with 1-dodecanol, a typical substrate for long-chain alcohol oxidases (LCAOs). Surprisingly, the enzyme from our screen displayed no sequence similarity to the known LCAOs. All previously characterised LCAOs belong to the glucose-methanol-choline (GMC) superfamily of oxidoreductases and are composed of an N-terminal FAD-binding domain containing a Rossmann fold, and a C-terminal substrate binding domain [117]. In contrast, the novel enzyme from our screen was predicted to contain an FAD-binding domain which spans the N- and C-termini of the protein, as well as a central substrate binding domain of a lactone oxidase (ALO)-type (Figure 13). The closest characterised enzymes to the archaeal LCAO, in terms of sequence identity, are L-gulonolactone dehydrogenase from *Mycobacterium tuberculosis* (27% sequence identity) and mouse L-gulonolactone oxidase (25% seq identity), which might explain why the protein was annotated as L-gulonolactone oxidase (EC 1.1.3.8). Although dissimilar in sequence, the enzyme displayed a typical activity profile of a LCAO, active with a range of fatty alcohols, with 1-dodecanol being the preferred substrate (Figure 13). Overall, our results confirmed that the archaeal enzyme is a novel type of LCAO, and together with the previously characterised PCMH-type LCAOs, they are an example of non-homologous isofunctional enzymes.

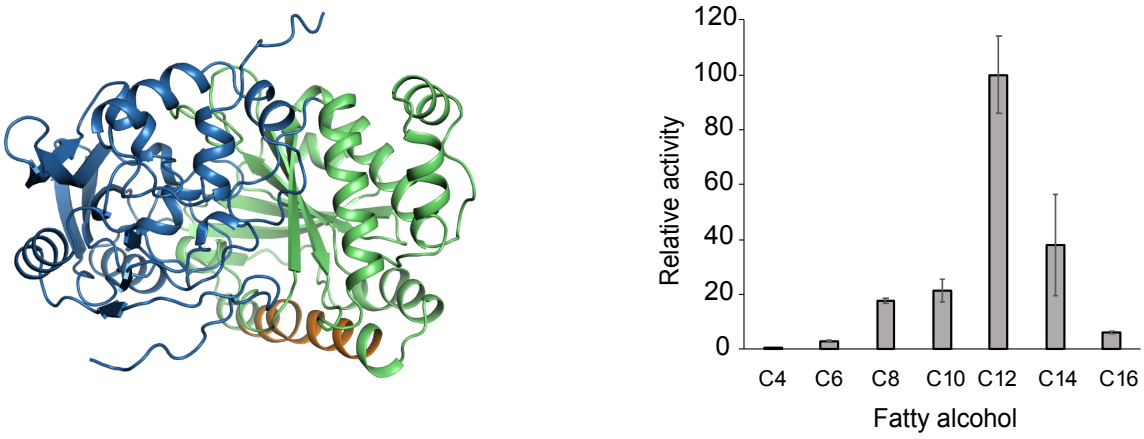

Figure 13. Characterisation of the archaeal long-chain alcohol oxidase. Left: structural model of LCAO. The PCMH-type FAD-binding domain is coloured in blue, the substrate binding domain is coloured in green, the membrane-bound helix is coloured in orange, the model obtained using AlphaFold Collab [118]. Right: Activity of LCAO with a range of fatty alcohols (C4: 1-butanol, C6: 1-hexanol, C8: 1-octanol, C10: 1-decanol, C12: 1-dodecanol, C14: 1-tetradecanol, C16: 1-hexadecanol). [111]

The results of Paper II show that our semi-untargeted approach of enzyme profiling proved useful in finding novel enzymes. The more classical screening used for enzyme families profiling might have been successful in finding the HexNAcO as well, since the enzyme is homologous to other known sugar oxidases. However, confirming activity for the novel archaeal LCAO would not be possible using targeted approaches, as the enzyme shows no similarity to the known alcohol oxidases. Our approach could also be potentially useful for detecting promiscuous activities of enzymes - side reactions that are distinct from the enzyme's main activity. At first thought to be of marginal relevance, they are now considered crucial for investigating protein evolution [119], as well as designing *in vitro* metabolic pathways [93].

Although successful in finding novel enzymes, our method has its limitations. Like many other high-throughput activity profiling platforms, we screened purified enzymes, and their recovery can be low, in our case 58%. Additionally, in the “all-vs-all” screen we focused on testing only the oxidation reaction, so any enzyme with a different type of activity would not be detected. Also, the type of substrates tested was limited to the ones commercially available, and their number limited to suit the format of a 384-well plate. Finally, it is worth noting that the selection process of the sequences for the “all-vs-all” experiment could have been improved by filtering out those sequences with no identity to the known EC 1.1.3.* representatives. Since the misannotated sequences made up a big part of the tested proteins, replacing them with the EC 1.1.3.* homologues would give a more accurate view of the enzyme class’ sequence space.

2.4.3. Biological function of proteins

“Protein function” can be defined in many ways, but the fullest description allows answering two questions: WHAT does the protein do and WHY. One caveat in many approaches of finding new enzymes, including ours, is focusing only on the molecular function of an enzyme - the WHAT. The biological function - the WHY - oftentimes remains unstudied. This is particularly true for activities from non-model organisms. Sometimes biological functions for an enzyme can be proposed by inferring them from other, better studied, organisms. This might be true for the archaeal LCAO characterised in Paper II, as a biological role for this enzyme in yeast is proposed to be involved in utilising fatty acids as an energy source [120]. It is possible that a similar pathway is present and utilised by marine archaea. However, the

biological function of the bacterial HexNAcO remains even more elusive, as no pathway directly utilising this enzyme has ever been described. Other peroxide-producing saccharide oxidoreductases are proposed to play a role in competing with other organisms through oxidative stress [121]. Since the closest homologues of HexNAcO from *R. solanacearum* are present in other soil and water inhabiting bacteria, it is possible that the enzyme plays a role in helping the host to thrive in those particular habitats.

Investigating biological functions is rarely the main focus of studies looking for enzymes with industrial applications. More care is put towards precise characterisation of enzymes' biochemical and biophysical properties or optimising them for industrial use. However, looking from the systems biology perspective, it is important to consider the biological questions as well. The majority of our current knowledge in biochemistry has been based on studying a handful of model organisms, yet the examples of enzymes from "niche" organisms revealed that the chemistry of life is much more diverse than expected [99]. In a similar manner, investigating novel activities from the biological perspective, although often very difficult, can perhaps change a lot about what we consider a "classical" biology.

Chapter 3. Investigating sequence space of engineered enzymes

3.1. Protein design

Discovery of novel enzymes is often guided by their potential for application. Enzymes as biocatalysts have become an important part of “green chemistry” - a design of more sustainable processes with minimized environmental impact [122]. They are being employed for plastic degradation [123], synthesis of pharmaceuticals [124], pest control [125], or processing plant biomass for biofuel, feed, food, and paper production [126]. However, as described in Chapter 2, looking for desired activities in nature is not an easy task. Even if an enzyme catalysing a reaction of choice is found in nature, it might lack properties necessary for an industrial setting, for example stability at high temperatures or low pH. Engineering enzymes for desired activities and properties seems like the most obvious solution for this problem, although not as straightforward as one can hope for. With the invention of site-directed mutagenesis, it became possible to study protein variants with changed residues [127]. It was a particularly useful tool for investigating the importance of catalytic residues. However, the attempts to introduce desirable properties with this technique were not hugely successful, as the effects of mutations were often unpredictable. A shift towards a more random approach, site-saturation mutagenesis, proved more successful [128]. In this approach, targeted residues are changed for all possible amino acids, rather than a specific one. Both methods, however, require a lot of knowledge about the enzyme of interest, like structure or identified catalytically important residues. What turned out to be the most successful approach in enzyme engineering, involved taking one step further into randomness. Directed enzyme evolution subjects a protein of interest to rounds of random mutagenesis and selection for a desired feature in the “fitness landscape” (Figure 14). Frances Arnold applied this method for the first time in 1993 to design a protease active in a nonnatural environment: high concentrations of an organic solvent [129]. DNA shuffling was introduced one year later by Willem Stemmer, as a tool mimicking DNA recombination to gain improved features [130]. It proved particularly useful in combination with directed evolution, allowing the creation of chimeras of several mutants. Successful attempts of enzyme evolution followed, allowing the creation of enzymes with enhanced activities in low or high temperatures, pH, salt concentrations, as well as activity toward non-natural substrates [6]. Enzyme evolution

techniques allowed not only for optimization of existing activities, but also creation of enzymes catalysing reactions yet undiscovered in nature, using promiscuous activities as a starting point [131]. Being an undeniably powerful tool for engineering, direction evolution can also be applied for fundamental research; to test evolutionary theories and reproduce evolutionary scenarios [132].

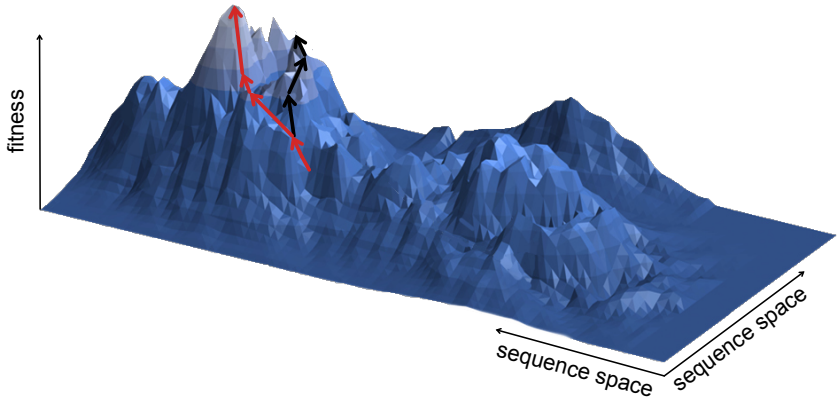

Figure 14. Directed evolution of proteins resembles climbing a “fitness landscape” mountain of a sequence space. Each round of selection results in choosing the mutant with the highest fitness with the aim to reach a fitness optimum (red line), although the presence of local optima might restrict some of the mutational paths uphill (black line). Adapted from [133].

3.2. Machine learning in protein design

One of the things learned from directed evolution studies, was that the beneficial mutations are often found in unexpected parts of enzymes - not only in the catalytic core. This explains why the early attempts of rational design were not as successful as it was hoped for. If human intelligence still struggles to understand and outsmart nature, perhaps artificial intelligence can? Machine learning (ML) is a branch of artificial intelligence focusing on data analysis to learn distribution of the data, make predictions, or generate new data. It is currently used for such applications as image or speech recognition and medical diagnosis [134]. In enzyme research, ML tools use biological data, such as protein sequence or structure to extract patterns [94]. These patterns are then used to classify new enzymes, predict their features, and find new variants with better catalytic properties [135]. Algorithms predicting optimal catalytic temperatures or solubility enable narrowing down variants with desired properties or increased chances of producing proteins heterologously [95,136–138]. Enzyme class predictors and substrate identifiers can be of help with functional annotation efforts, as well as guiding experimental design [139,140]. Much of the existing work on ML-guided enzyme

design is focused on incorporating discriminative models which predict properties of a given sequence by training on labelled sequence/structure-fitness pairs [141]. In contrast, generative models have a different kind of approach, by taking advantage of unlabelled protein sequences. Such methods can learn the underlying data distribution and produce new samples from it [141]. Paper III describes the ProteinGAN generative model, a machine learning approach which enables generation of novel functional enzymes with natural-like biochemical properties [142].

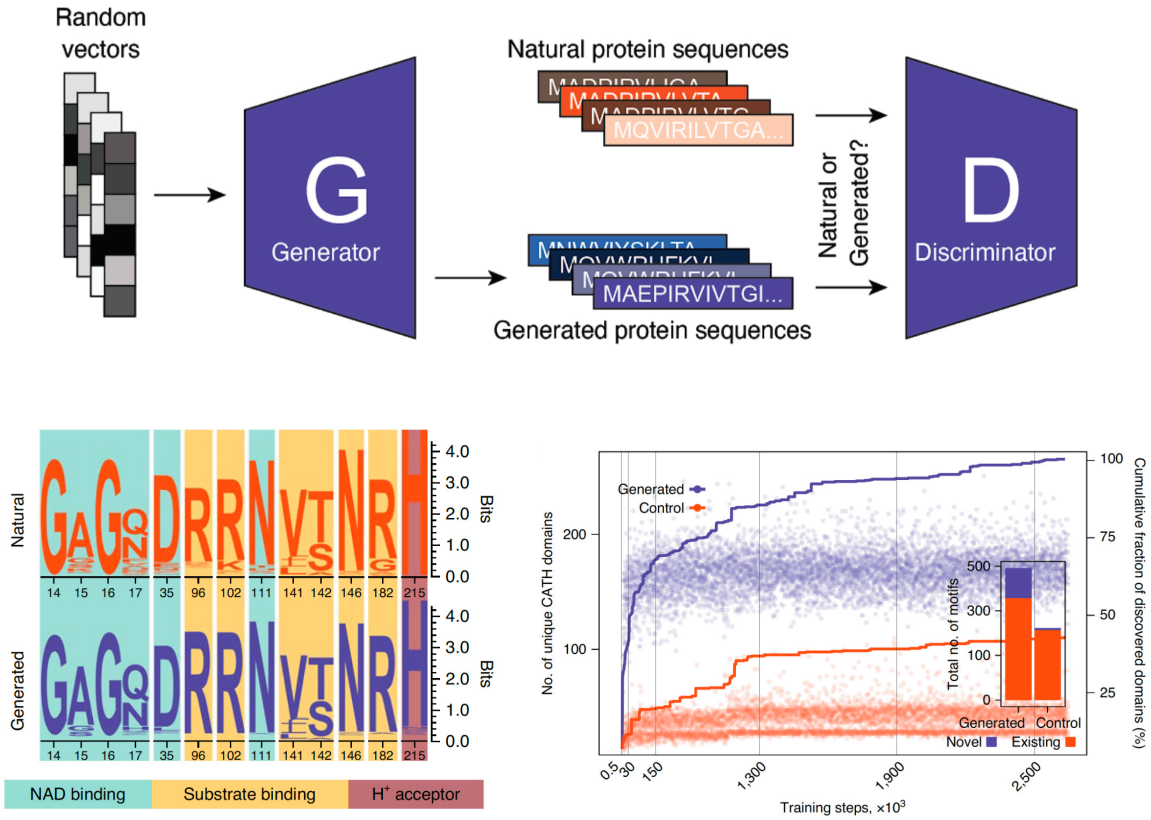

Figure 15. ProteinGAN mimics and expands the natural enzyme sequence space. Top: ProteinGAN training scheme. Given a random input vector, the Generator network produces a protein sequence, which is scored by the Discriminator network by comparing it to the natural protein sequences. The generator tries to fool the discriminator by generating sequences that will eventually look like real ones. Bottom, left: key conserved positions of natural MDH are also conserved in ProteinGAN-generated sequences. Bottom, right: CATH [143] domain diversity expanded throughout the evolution of ProteinGAN training, as opposed to a control of randomly mutated sequences from the training dataset. Insert: ProteinGAN generated novel domains that are not present in the existing MDH family, while random mutation causes decrease of diversity. Adapted from [142].

3.3. ProteinGAN (Paper III)

In order to mimic natural enzymes, the ProteinGAN method made use of generative adversarial networks (GAN) tailored to learn patterns from amino acid sequences. GAN is based on two neural networks, generator and discriminator, that compete with each other [144]. The generator network produces a protein sequence, which is then scored by the discriminator network by comparing it to known protein sequences from the training dataset (Figure 15). The neural network was trained on sequences of bacterial malate dehydrogenase (MDH, EC 1.1.1.37), a core enzyme of the tricarboxylic acid cycle, which catalyses the conversion of malate to oxaloacetate using NAD+ as a cofactor. Two phylogenetic groups of MDH are known, which display very low sequence similarity, but high structural similarity [145]. 20000 MDH-like sequences were generated using ProteinGAN, which were used to evaluate its performance. We showed that ProteinGAN was able to learn the patterns of natural MDH: preserve the key substrate-binding and catalytic residues (Figure 15), or local amino acid relationships. Additionally, ProteinGAN managed to expand the known MDH sequence space, for instance by producing novel structural motifs that do not exist in the training data (Figure 15). As a final verification step, the GAN-generated sequence space was experimentally tested for the MDH activity. 55 sequences with 45-98% sequence identity to natural MDH were sampled and synthesised. They were expressed and purified using the high-throughput experimental platform described in Chapter 2.1. The majority of the sequences expressed well but were not soluble (Figure 16). The enzymes with low sequence similarity to natural MDH were particularly prone to aggregation: no protein with less than 65% sequence identity to natural MDH was soluble. Out of the 19 soluble proteins, 13 displayed MDH activity, including one that shares only 66% identity with the closest existing enzyme.

The lack of activity of the remaining GAN-generated proteins might be partially explained by incorrect folding of the proteins, since it is known that even natural enzymes might be problematic to express and/or fold while heterologously produced in *E. coli* [146]. One example of issues with folding correctly comes from the GAN56 sequence, the only inactive enzyme that was both highly expressed and soluble. Melting profiles of the GAN-MDH proteins were investigated after publication of the Paper III results, and revealed that GAN56 profile is atypical - elongated, with two humps - overall very different from the melting profiles of the active GAN sequences or natural MDH (Figure 16). This might indicate that inactivity of GAN56 is caused by issues with folding, and it is hard to establish whether these are caused by the design or heterologous production. The melting profile screen of the GAN-generated

enzymes also revealed their wide range of thermal stability (Figure 16). Interestingly, ProteinGAN proved able to cause acquisition of thermostable features, which is visible on the example of the sequences GAN35 and GAN29. They share a 90% sequence identity, and a common closest natural sequence, from a mesophile *Brucella ceti*, yet they display a 12°C difference in the melting temperatures (Figure 16).

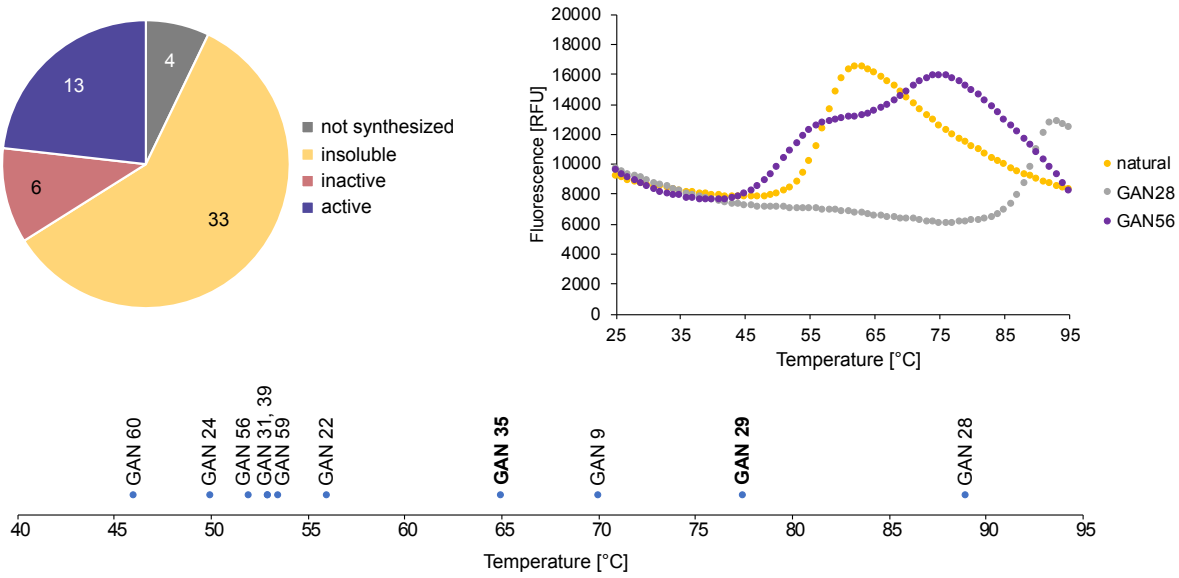

Figure 16. Experimental characterisation of ProteinGAN-generated MDH sequences. Top, left: summary of experimental results of the 56 generated MDH sequences. Top, right: melting profiles of selected ProteinGAN-generated MDHs obtained by a thermal shift assay. Bottom: melting temperatures of selected ProteinGAN-generated MDHs.

Despite the growing number of theoretical approaches for biological sequence generation, until recently their ability to generate novel functional proteins was limited. The results of the Paper III prove that it is possible to mimic nature by generating enzymes that behave like natural ones. In the study, ProteinGAN was able to produce functional proteins with up to 106 amino acid substitutions in comparison to the closest natural enzyme - not a trivial number, considering that around a third of single amino-acid substitutions result in a loss of protein's function [147]. Rational design of proteins has no such comparable success, simply because we still struggle to understand the sequence-function relationship of proteins. The ProteinGAN model, by learning the underlying sequence data distribution, was able to bypass this gap in our knowledge. Around the same time as Paper III, other studies also confirmed protein design capabilities of machine learning, using different deep generative algorithms and model proteins [148,149], although like our study, they mostly focus on generation of natural-like sequences. However, the ultimate goal of a protein engineer is to improve, not reproduce. ProteinGAN was able to introduce thermal stability to sequences (Figure 16), thus acquisition

of a desired function is viable. Additionally, the GAN-designed sequence space expanded into novel structural domains (Figure 15). Although these were not included in the experimentally tested set, it shows that expansion of structural (and potentially functional) diversity is possible.

3.4. ML-enabled protein design: promises and challenges

Several studies have already shown promising results of protein optimization by generative models, although this direction is still in its infancy [141]. As work on engineering enzymes by generative models progresses, successful attempts of ML enzyme design have been made using mostly discriminative approaches [141]. In ML-guided directed evolution, mutant sequence space is sampled by an algorithm to choose promising candidates for screening, thus minimizing the experimental part [150]. ML algorithms can screen much larger swaths of sequence space than any ultrahigh-throughput assay and are additionally able to escape local activity optima by learning about the entire functional landscape. However, when planning a ML-guided directed evolution project, one must take into consideration the costs and time involved in analysis and sequencing of the training dataset, and synthesis of the predicted high performing variants. It might be a particularly good option for proteins with an expensive or low-throughput activity screen, as shown for a light-gated channelrhodopsin, for which a high light sensitivity variant was obtained using only 102 proteins in a training dataset [151]. In addition to optimizing single proteins, discriminative ML algorithms have been used to engineer whole biosynthetic pathways. Combined with genome scale models, ML allows to pinpoint engineering targets and can result in flux optimisation [152] or improved production of desired metabolites [153,154].

All ML algorithms, regardless of the type, rely on data to perform predictions or generate new variants. In case of ML-enabled enzyme design, this data is usually protein sequence, structure, and information connecting sequences to their properties. For the best results, this data needs to be of good quality, plentiful, and unbiased. This, however, is not the reality scientists working with ML are facing at the moment. Available protein databases provide invaluable resources for training ML algorithms, however each of them comes with certain caveats [94]. Sequence databases, such as TrEMBL, contain a large number of entries, although the majority is not connected to a known function or feature. Additionally, as described in Chapter 2.2. of the thesis, some annotations in such databases are far from trustworthy. In Paper III, the training dataset consisted of bacterial sequences annotated as

MDH in UniProt, without pre-filtering them for their similarity to characterised sequences. Fortunately, the vast majority of the sequences displayed high sequence identity to the curated sequences (Figure 17), but as discussed in Paper I, this is not the case for many enzyme classes. Curated databases, like Swiss-Prot, offer high quality datasets, although their size is limited. Functional databases, like BRENDA, contain a comprehensive overview of enzyme literature, but as the data come from many different sources, they are inevitably inhomogeneous. Overall, the majority of data containing enzyme characterisation are not collected with a ML application in mind. As already discussed in Chapter 2.2 of the thesis, systematic registration of experimental data is a big issue, and a lot of effort in collecting training data might be put into their extraction and cleaning. A set of “FAIR” guiding principles for scientific data management and stewardship was recently published to improve data Findability, Accessibility, Interoperability, and Reuse [155]. Initiatives such as STRENDA, aiming to provide guidelines of standards for reporting enzymology data, could help to solve the problem of inhomogeneous enzyme data, but only if the whole scientific community complies [156].

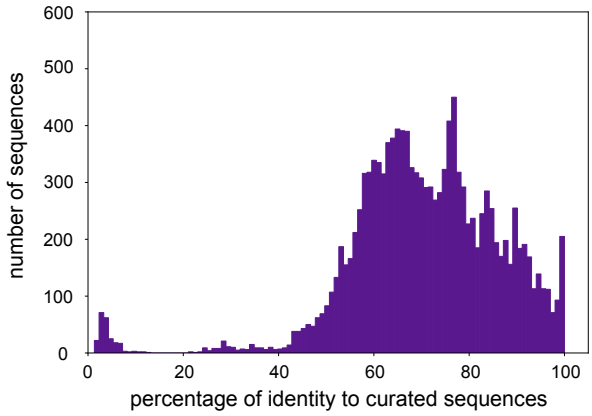

Figure 17. Distribution of sequence identities of the ProteinGAN training dataset to the closest curated natural sequence (present in Swiss-Prot).

The issues with experimental data compatibility with ML do not only concern the quality or reporting. As discussed at the beginning of Chapter 2 of the thesis, experimental characterisation is often heavily biased towards model organisms, or organisms with industrial or medical relevance. Additionally, negative results are very rarely reported. As the ML-based predictors usually require uniform sampling of data, these biases inevitably affect performance of the models. Advances in next generation sequencing are helping to alleviate this bias, as more and more sequences are being deposited to public databases, including those from

uncultured microbial species (Figure 4). Experimental platforms focusing on uniform investigation of sequence space might help to resolve the problems of biased selection of sequences for characterisation, as well as the lack of negative data reporting. Ultrahigh-throughput methods for enzyme screening applied in directed evolution or protein discovery can enable investigation of even larger sequence spaces and provide more data for the ML training sets.

Chapter 4. Enzyme assays for sequence space investigation

4.1. Enzyme assays

To get a full understanding of how an enzyme works, it is necessary to study the reaction it catalyses. Enzyme assays are laboratory tools that enable investigation of enzymatic reactions by measuring either the consumption of substrates or production of products. The first ever studies on enzymatic activities were conducted on digestive enzymes, as it was possible to observe the results of the enzyme action with the naked eye: breakdown of grains or post-mortem lesions of stomach walls [10]. With time, more advanced methods appeared that enabled identification, as well as quantification of substrates and products of enzymatic reactions [157]. These methods can be divided into direct and indirect. Direct methods measure change in properties of a reaction mixture, such as change in viscosity, colour, light polarization, absorbance, fluorescence, or chemiluminescence. For example, Menten and Michaelis when investigating the kinetics of enzyme invertase, which digests sucrose to glucose and fructose, relied on a difference in light polarisation properties between the product and substrates [15]. Another classical example of a direct enzyme assay is applied when studying NAD-dependent enzymes, as NADH displays an absorbance peak at 340 nm, while its oxidised form, NAD+, has virtually no absorbance at this wavelength. A different approach for a direct detection of products can also be performed based on analysis of their absorption and mass spectra using high-performance liquid chromatography or mass spectrometry [158,159]. In indirect methods, one of the products of the reaction of interest is used as a substrate of another, easily detectable reaction. The second reaction can be either enzymatic, or non-enzymatic. Such coupled enzyme assay was for instance used in Paper I to study S-2-hydroxyacid oxidase activity: hydrogen peroxide, a reaction product, is used as a substrate of horseradish peroxidase which oxidises a non-fluorescent substrate into a fluorescent product (Figure 8). Labelling of substrates, either with radiolabelling or using fluorescent probes, is also used to investigate enzymatic reactions that cannot be easily measured directly or via coupled assays. Detection of ligand binding to an enzyme can also be another method of studying enzymes, without the direct detection of substrates or products.

Selection and optimisation of an enzyme assay is a crucial step in enzyme research [157,160]. The choice of an assay depends not only on the type of the reaction catalysed, but also on the final goal of the enzyme study, costs, levels of details and throughput required. Enzyme discovery and directed evolution approaches require a particularly high throughput screening setup. In the original article describing the first attempt of directed evolution, only 300 mutated variants were screened [161]. Nowadays, screening of thousands to millions of variants is performed in search of the best performers. The ultrahigh-throughput platforms (up to 10^8 screened variants per day) rely on compartmentalization of reaction components in cells, synthetic droplets, or microchambers, screening for the desired activity, and library sorting [162,163]. In such setups, fluorescence-based assays are applied, and screening is performed using fluorescent-activated cell sorting, microfluidics, or fluorescent microscopy [163]. Oftentimes availability of a high-throughput screening method is a bottleneck for engineering a specific enzyme, although new methods for fluorescent and non-fluorescent screening and sorting are being developed [164,165]. Microplate-based assays are commonly used for drug screening, sequence-space investigations of enzyme families, and can also be applied in directed evolution. They rely on assays performed in multiwell plates (96 -1536 wells), in which both fluorescence and absorbance of the reaction mixture can be screened. Although their throughput is much lower than the ultrahigh-throughput platforms, it can be improved using liquid handling robots, up to the 10^6 variants screened per day [94,166].

4.2. Kinetic enzyme assays

High-throughput platforms for enzyme discovery, engineering, and sequence space screening usually produce a single readout per variant, providing only the information on increased or decreased activity. For more thorough characterisation, kinetic assays are used. In such setups, reaction rate is the crucial value being measured - a concentration of substrate disappearing, or product produced as a function of time. Enzyme kinetic studies enable investigation of the catalytic mechanism of an enzyme, inhibition and activation, or comparison of its activity profile to similar enzymes. Most commonly, enzyme kinetic analysis is performed in the steady-state, where concentration of the enzyme-substrate complex remains constant. Historically, the discontinuous assays were used for enzyme kinetic studies: the samples are removed at intervals from the reaction mixture, and the amount of product formed, or substrate consumed, is calculated. Nowadays, the most prevalent routines adopt continuous assays, in which a signal can be recorded periodically over time. Although many improvements have

been made over time, traditional kinetic assays remain time- and reagent-consuming, as well as low-throughput, which usually results in testing a limited number of substrates and conditions. As already stated in previous chapters of the thesis, there is a great need for more enzyme-related data. While most high-throughput screening platforms focus on providing endpoint measurements, development of high-throughput platforms providing kinetic data is still in its infancy. In Paper IV we adapted a qPCR microfluidic system to perform enzyme kinetic measurements with improved throughput and decreased sample usage compared to classical multi-well assays [167].

4.3. Adaptation of a commercial qPCR platform for enzyme kinetic studies (Paper IV)

Commercial microfluidic qPCR platforms enable automated gene expression analysis at the nanoliter scale. One such platform was developed by Fluidigm Corporation and relies on using integrated fluidic circuit chips [168]. Samples and reagents are first pippeted on a chip and pressure-loaded into its reaction chambers. The chip is then transferred to a real-time PCR machine which performs thermal cycling and images the chip in real time. The Fluidigm system offers a wide range of chips with different reaction chamber setups, numbers, and volumes. To test the usability of the Fluidigm platform for enzyme screening, we used the most basic chip, FlexSix, containing six partitions, each with 12 wells on the assay and sample sides (Figure 18). It offers a medium-throughput range, from 144 up to 864 reactions per chip. We used lactate oxidase as a model enzyme, and a hydrogen peroxide detecting fluorescent assay, which is based on a principle described in Chapter 2.1 (Figure 8). Preliminary results showed that activity of the enzyme can be detected in the microfluidic system, and its initial reaction rate was successfully recorded (Figure 18). To evaluate if the system is suitable for obtaining kinetic parameters of enzymes, we measured initial reaction rates at different substrate concentrations for three peroxide-producing oxidases: lactate oxidase, glucose oxidase, and glutamate oxidase. The obtained kinetic values were reproducible, and comparable to those obtained using a standard setup in a 384-well plate (Table 2). In comparison to the standard method, the Fluidigm platform offered a 2000-fold decrease of reaction volumes - from 20 µl to 8.9 nl. It also involved less manual handling, as the mixing of enzyme and substrate happened on the chip. Using only half the capacity of the most basic

chip, the tested setup allowed for establishing kinetic parameters for three different enzymes in one run (432 reactions), while three separate runs had to be performed using microplates.

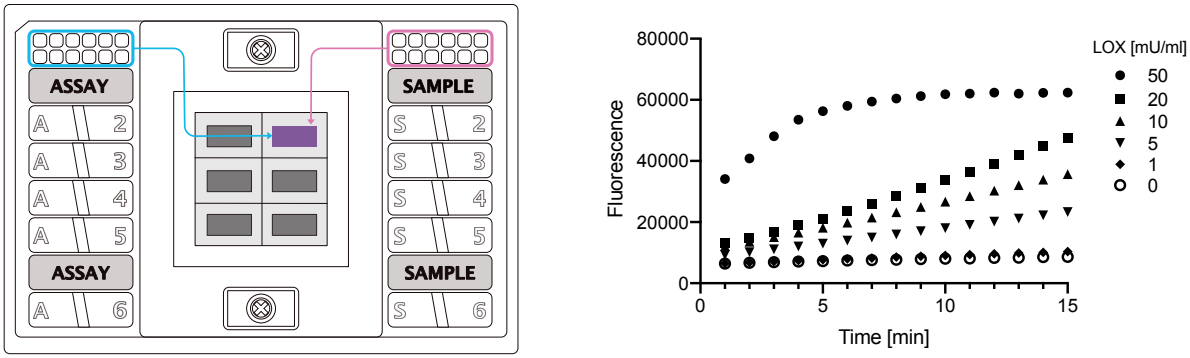

Figure 18. Fluidigm qPCR platform for enzyme assays. Left: schematic representation of a FlexSix microfluidic chip. Solutions from 12 assay inlets are mixed with solutions from 12 sample inlets in a 1:9 ratio. The final reaction volume is 8.9 nl. Right: example of the obtained initial reaction rates of lactate oxidase in reaction with lactate. [167]

Table 2. Comparison of kinetic values of three oxidases obtained in the Fluidigm system and a microplate reader. [167]

Enzyme	System	K_M [mM][a]	V_{max} [µmol mg^{-1} min^{-1}][a]
Lactate oxidase	microplate	1.62 ± 0.48	22.00 ± 1.15
	chip	1.23 ± 0.12	23.67 ± 2.73
	literature[b]	0.5 - 1	114 - 270
Glutamate oxidase	microplate	0.20 ± 0.01	6.30 ± 0.25
	chip	0.29 ± 0.08	7.30 ± 1.00
	literature[b]	0.17 - 0.3	6 - 55
Glucose oxidase	microplate	20.67 ± 2.72	2.40 ± 0.00
	chip	16.33 ± 1.76	23.00 ± 4.58
	literature[b]	22 - 32	6 - 170

[a] - Values for microplate and chip represent mean average (± standard error of mean; n = 3).

[b] - Literature values range as reported for wild type enzymes in BRENDA DB [169].

The use of microfluidic devices for enzyme kinetic studies is not a novel idea; development of such systems is becoming increasingly common [170]. One of the most rapidly evolving systems are droplet-based microfluidics devices, which rely on a substrate being mixed with an enzyme into nanoliter droplets. Such droplets can later be tracked, and activity signals can be recorded over time. A great advantage of the approach is a possibility of very rapid mixing of reagents, allowing to probe pre-steady states of enzymes. Additionally, the setup allows for

the creation of concentration gradients, resulting in greater precision than in the standard plate methods or the Fluidigm system. Recently one such platform has been used to gain more insight into kinetics and thermodynamics of native and engineered enzyme variants of haloalkane dehalogenase [171]. However, the vast majority of published microfluidic methods for studying enzyme kinetics are proof-of-concept, displaying new methodologies, rather than being applied for answering real-life scientific questions [170]. The lack of easy to use, standardised systems has been discussed as a major cause of low adaptation of microfluidics by nonspecialists such as biochemists [172]. The Fluidigm system offers a commercial alternative for high-throughput kinetic screening of enzymes, requiring no knowledge of how to run and collect data with a microfluidic device. Its components are easy to operate, and data can be collected using intuitive software. Data analysis can be performed by the R scripts developed by us or with the use of other available tools [173]. A big potential of the Fluidigm system is the possibility of parallelization of many different kinetic measurements and testing of different enzymes, substrates and conditions in a single run. With many different chips available, users can adapt the layout and throughput to their needs: from smaller scale pilot studies to large scale, all-vs-all type of studies, where 9216 reactions can be run simultaneously. Certainly, the setup has its downsides too. Unlike some current microfluidic platforms, the Fluidigm system does not allow rapid probing of reactions, and manual preparation of substrate concentrations is still necessary. As such, it resembles more of a standard plate-based assay in a miniaturised version. Additionally, like many high-throughput commercial systems, its acquisition requires considerable investments. However, high-throughput qPCR systems are becoming increasingly popular, and many genomics core facilities are being equipped with such devices, offering access to them for the scientific community.

Novel methods and platforms for enzyme research are constantly being developed. By adapting a microfluidic qPCR device for enzyme research, we showed an alternative path of method development: repurposing an already developed scientific methodology to a different use. Such an approach has been a crucial part of science for many years. Early enzyme scientists adapted methods for measurements of biophysical and chemical properties of solutions to develop enzyme assays. A setup of a standard qPCR machine was used to develop a thermofluor assay for measuring protein melting temperature [174]. The Fluidigm device itself was adopted to develop a successful commercial protein biomarker discovery platform [175]. Although the development of new scientific tools shall not (and certainly will

not) cease, repurposing of already existing methods should also be considered as a potential option in method development.

Experimental methods for investigating the activity of enzymes are crucial for discovery of new proteins, engineering novel variants, improving functional annotations, understanding functions on molecular levels, and many others. No single approach will cover all the needs, and development of many different types of methods for enzyme investigation is necessary: those allowing for precise kinetic investigations, those for ultrahigh-throughput screening, and all the methods in between. Applied together with other methods for studying proteins, as well as bioinformatic approaches, they will allow us to learn more about the sequence space of enzymes.

Chapter 5. Outlook

In 1955 in his “crystal ball” lecture about the future of enzyme research, Linus Pauling said that “when we understand enzymes - their structure, the mechanism of their synthesis, the mechanism of their action - we shall understand life” [176]. With time we found (at least some) answers to those questions, yet it would be bold to state we understand life. Rather, we managed to unveil a tiny bit of the mystery, only to see there is much more to understand. And yet, undoubtedly, enzymes are one of the central molecules that make up life. That is one of the reasons why they remain a critical subject of scientific studies: as a driving force of metabolism, focal points of genetic disorders, or drug targets. At the same time, their catalytic potential is being explored for industrial applications and processes.

There are many layers to the study of enzymes: finding new activities, investigating their mechanisms, properties, interactions, side activities. A crucial part of protein studies is also the development of new scientific tools. All these layers are connected and influence each other greatly. Similarly, basic research is a constant source of inspiration for more industrially aimed research, and vice versa. The work presented in the four papers and discussed by me in this thesis focused on investigations of larger swaths of enzymatic sequence space. The starting point of my work was to explore the natural diversity of enzymes, and with time it has meandered across the topics of enzyme discovery, annotation, design, as well as assay platform development. Taking its inspiration from both basic and applied research, hopefully my work contributed on some level to both areas. Similar large-scale protein profiling attempts will most likely become ever more common. Although it is hard to predict what the future of protein research holds, the developments of novel experimental and modelling approaches will undoubtedly continue to move the field forward.

The advancement of investigative methods and tools have been crucial for modern science. Their development not only provided answers for existing questions, but most importantly opened doors for new discoveries. Methods developed throughout the 20th century, like protein purification and crystallization, molecular cloning, or PCR, allowed many scientific fields to move forward, including the field of enzyme research. We were able to take a closer look at how enzymes are built and how they work. In the 21st century, with the development of affordable DNA sequencing and synthesis methods, high-throughput approaches appeared, allowing for investigating many sequences at the same time, including those from

unculturable organisms, or the ones not present in nature at all. Would the next step in enzyme research be to marry high-throughput with high-depth? Promising methods are already appearing, for instance a platform developed by the groups of Polly Fordyce and Dan Herschlag, which does the incredible work of expression, purification, and biochemical characterisation of over one thousand enzyme variants, all in a one small chip. [177] Originally used to study the effects of single mutations, the platform could easily be applied for large-scale activity profiling of enzyme families or AI-designed sequences. With the use of such platforms, mechanistic investigations of enzymes, their phylogenetic analysis, characterization of allelic variants, or side activities could soon be performed in a high-throughput manner, at the same time with much more depth than ever.

Novel high-throughput tools to study enzyme activity generate a novel type of data. Right now, they are primarily used as an intermediate step for discoveries, such as novel activities or better performing variants, where at the end only selected candidates are comprehensively characterised. Yet, the “intermediate” data could deliver a lot more information apart from being a stepping stone to answering the original questions. For instance, large scale activity profiling could not only lead to discovery of novel enzyme families, but also be used as a starting point of evolutionary investigations or to guide functional annotations. As negative data make up a vital part in such datasets, they could also be reused for machine learning training purposes. However, for such data to be reused, a structured way of reporting and storing them is necessary. When asked about integrating results from medium/high-throughput enzyme characterisation studies into their database, UniProt representatives admitted that such studies provide meaningful biological insights, however, the database currently does not operate a systematic pipeline to specifically search for or integrate them. Would standards and databases for reporting such data be set up, similarly to those functioning for omics or compound screening datasets? [178,179] They most certainly should, although it would be far from trivial, considering we still struggle to report data from “low-throughput” experiments. Many articles discussing protein discovery, annotation or design conclude with a call for more data. However, if the data is not easily available and recorded in public databases, what is the point of producing it?

One of the most “data hungry” tools showing great promise in biology is machine learning. Currently trained mostly on the abundant protein sequence data, it could only improve with the addition of the extra layer of good quality biological information. A great example of how

powerful machine learning can be when applied to biological data, is a recent breakthrough in protein structure prediction: the AlphaFold 2 model developed by Google's DeepMind. [118,180] In November 2020 the team behind AlphaFold 2 won the 14th Critical Assessment of Structural Prediction competition, providing structural models with the accuracy that has never been seen before. Not long after, DeepMind launched a database with predictions of over 350000 structures from human and 20 model organisms, promising to release structures for many more proteins in the future. What is more, AlphaFold's code has been made open source, allowing scientists around the world to obtain structural models of any protein sequence. While its full potential and limitations are yet to be unraveled, AlphaFold would certainly be a crucial tool in many fields, including those of drug discovery, protein design, and structure-based functional annotation.

The progress of molecular biology, including the study of enzymes, has been enormous over the past decades; to think that a mere one hundred years ago we still did not know that DNA is the carrier of genetic information, or that enzymes are proteins. Although groundbreaking discoveries have been made since that time, many more questions are still waiting to be answered (and, perhaps more importantly, to be asked). One thing that can certainly be learned from the history of science, is to be mindful of and accept the biases in our knowledge, which have their roots in the available technologies, values or interests of current societies. The direction of enzyme research will certainly be shaped by those factors, most likely focusing on discovery and design of novel biocatalysts for industrial applications. Hopefully, new technologies will also bring benefit to fundamental research, perhaps shifting our human-centric view of biochemistry. For a long time, such fundamental questions have been conceived and answered in the minds of academics. However, some of the cutting-edge basic research is starting to emerge from industrial R&D teams. What would it mean for academia? Certainly big changes, the direction of which only time will show.

References

1. Fukao T, Nakamura K. Advances in inborn errors of metabolism. J Hum Genet. 2019;64: 65.

2. Clardy J, Fischbach MA, Currie CR. The natural history of antibiotics. Curr Biol. 2009;19: R437–41.

3. Kang TS, Georgieva D, Genov N, Murakami MT, Sinha M, Kumar RP, et al. Enzymatic toxins from snake venom: structural characterization and mechanism of catalysis. FEBS J. 2011;278: 4544–4576.

4. Littlechild JA. Enzymes from Extreme Environments and Their Industrial Applications. Front Bioeng Biotechnol. 2015;3: 161.

5. Singh R, Kumar M, Mittal A, Mehta PK. Microbial enzymes: industrial progress in 21st century. 3 Biotech. 2016;6: 174.

6. Heckmann CM, Paradisi F. Looking Back: A Short History of the Discovery of Enzymes and How They Became Powerful Chemical Tools. ChemCatChem. 2020;12: 6082–6102.

7. McGovern PE, Zhang J, Tang J, Zhang Z, Hall GR, Moreau RA, et al. Fermented beverages of pre- and proto-historic China. Proc Natl Acad Sci U S A. 2004;101: 17593–17598.

8. Kindstedt PS. The history of cheese. Global Cheesemaking Technology. Chichester, UK: John Wiley & Sons, Ltd; 2017. pp. 1–19.

9. Harari YN. Sapiens – A brief history of humankind. Random House; 2014 [cited 19 Aug 2021]. Available: https://hallowesbrown.net/Sapiens%E2%80%93a_brief_history_of_humankind.pdf

10. Williams HS, Williams EH. A History of Science: in Five Volumes. Harper; 1904.

11. Kohler RE Jr. The enzyme theory and the origin of biochemistry. Isis. 1973;64: 181–196.

12. Northrop JH. Nobel lecture: The Preparation of Pure Enzymes and Virus Proteins. https://www.nobelprize.org/prizes/chemistry/1946/northrop/lecture/. 1946. Available: https://www.nobelprize.org/uploads/2018/06/northrop-lecture.pdf

13. Fischer E. Einfluss der Configuration auf die Wirkung der Enzyme. Ber Dtsch Chem Ges. 1894;27: 2985–2993.

14. Koshland DE. Application of a Theory of Enzyme Specificity to Protein Synthesis. Proc Natl Acad Sci U S A. 1958;44: 98–104.

15. Michaelis L, Menten ML, Johnson KA, Goody RS. The original Michaelis constant: translation of the 1913 Michaelis-Menten paper. Biochemistry. 2011;50: 8264–8269.

16. Blow D. So do we understand how enzymes work? Structure. 2000;8: R77–81.

17. Simoni RD, Hill RH, Vaughan M. Urease, the first crystalline enzyme and the proof that enzymes are proteins: the work of James B. Sumner. J Biol Chem. 2002;277: 23e.

18. Mulder GJ. Sur la composition de quelques substances animales. Bulletin des sciences physiques et naturelles en Neerlande. 1838;104: 9.

19. Hausmann R. To Grasp the Essence of Life: A History of Molecular Biology. Springer Science & Business Media; 2002.

20. Campbell ID. Timeline: the march of structural biology. Nat Rev Mol Cell Biol. 2002;3: 377–381.

21. Enzyme Commission, Historical Introduction. [cited 19 Aug 2021]. Available: https://www.qmul.ac.uk/sbcs/iubmb/enzyme/history.html

22. ExplorEnz: Enzyme Count. [cited 19 Aug 2021]. Available: https://www.enzyme-database.org/stats.php

23. Overturf K. Molecular Research in Aquaculture. John Wiley & Sons; 2009.

24. Hershey AD, Chase M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. J Gen Physiol. 1952;36: 39–56.

25. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature. 1953;171: 737–738.

26. Crick FHC. On protein synthesis. InSymposia of the Society for Experimental Biology; Number XII: The Biological Replication of Macromolecules. Cambridge University Press: Cambridge, UK; 1958.

27. Marshall RE, Caskey CT, Nirenberg M. Fine structure of RNA codewords recognized by bacterial, amphibian, and mammalian transfer RNA. Science. 1967;155: 820–826.

28. Jackson DA, Symons RH, Berg P. Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. Proc Natl Acad Sci U S A. 1972;69: 2904–2909.

29. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74: 5463–5467.

30. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2016;44: D67–72.

31. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409: 860–921.

32. Hayden EC. Technology: The $1,000 genome. Nature. 2014;507: 294–295.

33. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17: 333–351.

34. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020;21: 30.

35. DNA Sequencing Costs: Data. [cited 12 Oct 2021]. Available: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

36. JGI GOLD. [cited 12 Oct 2021]. Available: https://gold.jgi.doe.gov/statistics

37. Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. Q Rev Biophys. 2003;36: 307–340.

38. Rittié L, Perbal B. Enzymes used in molecular biology: a useful guide. J Cell Commun Signal. 2008;2: 25–45.

39. Novozymes. Enzymes at work. 2013.

40. Demirjian DC, Morís-Varas F, Cassidy CS. Enzymes from extremophiles. Curr Opin Chem Biol. 2001;5: 144–151.

41. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, et al. The PROSITE database. Nucleic Acids Res. 2006;34: D227–30.

42. Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol. 2016;17: 184.

43. Ellens KW, Christian N, Singh C, Satagopam VP, May P, Linster CL. Confronting the catalytic

dark matter encoded by sequenced genomes. Nucleic Acids Res. 2017;45: 11495–11514.

44. Gilks WR, Audit B, de Angelis D, Tsoka S, Ouzounis CA. Percolation of annotation errors through hierarchically structured protein sequence databases. Math Biosci. 2005;193: 223–234.

45. Goh KM, Shahar S, Chan K-G, Chong CS, Amran SI, Sani MH, et al. Current Status and Potential Applications of Underexplored Prokaryotes. Microorganisms. 2019;7. doi:10.3390/microorganisms7100468

46. Vickers CJ, Fraga D, Patrick WM. Quantifying the taxonomic bias in enzymology. Protein Sci. 2021;30: 914–921.

47. Rembeza E, Engqvist MKM. Experimental and computational investigation of enzyme functional annotations uncovers misannotation in the EC 1.1.3.15 enzyme class. PLoS Comput Biol. 2021;17: e1009446.

48. Bastard K, Smith AAT, Vergne-Vaxelaire C, Perret A, Zaparucha A, De Melo-Minardi R, et al. Revealing the hidden functional diversity of an enzyme family. Nat Chem Biol. 2014;10: 42–49.

49. Zhang X, Carter MS, Vetting MW, San Francisco B, Zhao S, Al-Obaidi NF, et al. Assignment of function to a domain of unknown function: DUF1537 is a new kinase family in catabolic pathways for acid sugars. Proc Natl Acad Sci U S A. 2016;113: E4161–9.

50. Baier F, Tokuriki N. Connectivity between catalytic landscapes of the metallo-β-lactamase superfamily. J Mol Biol. 2014;426: 2442–2456.

51. Mashiyama ST, Malabanan MM, Akiva E, Bhosle R, Branch MC, Hillerich B, et al. Large-scale determination of sequence, structure, and function relationships in cytosolic glutathione transferases across the biosphere. PLoS Biol. 2014;12: e1001843.

52. Huang H, Pandya C, Liu C, Al-Obaidi NF, Wang M, Zheng L, et al. Panoramic view of a superfamily of phosphatases through substrate profiling. Proc Natl Acad Sci U S A. 2015;112: E1974–83.

53. Vanacek P, Sebestova E, Babkova P, Bidmanova S, Daniel L, Dvorak P, et al. Exploration of Enzyme Diversity by Integrating Bioinformatics with Expression Analysis and Biochemical Characterization. ACS Catal. 2018;8: 2402–2412.

54. Helbert W, Poulet L, Drouillard S, Mathieu S, Loiodice M, Couturier M, et al. Discovery of novel carbohydrate-active enzymes through the rational exploration of the protein sequences space. Proc Natl Acad Sci U S A. 2019;116: 6063–6068.

55. Caparco AA, Pelletier E, Petit JL, Jouenne A, Bommarius BR, Berardinis V, et al. Metagenomic mining for Amine dehydrogenase discovery. Adv Synth Catal. 2020;362: 2427–2436.

56. Bastard K, Perret A, Mariage A, Bessonnet T, Pinet-Turpault A, Petit J-L, et al. Parallel evolution of non-homologous isofunctional enzymes in methionine biosynthesis. Nat Chem Biol. 2017;13: 858–866.

57. Linster CL, Van Schaftingen E. Vitamin C. Biosynthesis, recycling and degradation in mammals. FEBS J. 2007;274: 1–22.

58. Esser C, Kuhn A, Groth G, Lercher MJ, Maurino VG. Plant and animal glycolate oxidases have a common eukaryotic ancestor and convergently duplicated to evolve long-chain 2-hydroxy acid oxidases. Mol Biol Evol. 2014;31: 1089–1101.

59. Su J, Hirji R, Zhang L, He C, Selvaraj G, Wu R. Evaluation of the stress-inducible production of choline oxidase in transgenic rice as a strategy for producing the stress-protectant glycine betaine. J Exp Bot. 2006;57: 1129–1135.

60. Hubbard BK, Thomas MG, Walsh CT. Biosynthesis of L-p-hydroxyphenylglycine, a non-proteinogenic amino acid constituent of peptide antibiotics. Chem Biol. 2000;7: 931–942.

61. Clausnitzer D, Piepersberg W, Wehmeier UF. The oxidoreductases LivQ and NeoQ are responsible for the different 6'-modifications in the aminoglycosides lividomycin and neomycin. J Appl Microbiol. 2011;111: 642–651.

62. Katayama K, Kobayashi T, Oikawa H, Honma M, Ichihara A. Enzymatic activity and partial purification of solanapyrone synthase: first enzyme catalyzing Diels-Alder reaction. Biochim Biophys Acta. 1998;1384: 387–395.

63. Metkar SK, Girigoswami K. Diagnostic biosensors in medicine – A review. Biocatal Agric Biotechnol. 2019;17: 271–283.

64. Thakur MS, Ragavan KV. Biosensors in food processing. J Food Sci Technol. 2013;50: 625–641.

65. Sheldon RA, Brady D, Bode ML. The Hitchhiker's guide to biocatalysis: recent advances in the use of enzymes in organic synthesis. Chem Sci. 2020;11: 2587–2605.

66. Bell EL, Finnigan W, France SP, Green AP, Hayes MA, Hepworth LJ, et al. Biocatalysis. Nature Reviews Methods Primers. 2021;1: 1–21.

67. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28: 235–242.

68. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47: D506–D515.

69. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. Nucleic Acids Res. 1996;24: 21–25.

70. Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, et al. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res. 2021;49: D498–D508.

71. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44: D457–62.

72. Karp PD, Riley M, Paley SM, Pellegrini-Toole A. The MetaCyc Database. Nucleic Acids Res. 2002;30: 59–61.

73. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014;42: D490–5.

74. Passardi F, Theiler G, Zamocky M, Cosio C, Rouhier N, Teixera F, et al. PeroxiBase: The peroxidase database. Phytochemistry. 2007;68: 1605–1611.

75. Salzberg SL. Next-generation genome annotation: we still struggle to get it right. Genome Biol. 2019;20: 92.

76. Kyrpides NC, Ouzounis CA. Whole-genome sequence annotation: "Going wrong with confidence." Mol Microbiol. 1999;32: 886–887.

77. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol. 2009;5: e1000605.

78. Devos D, Valencia A. Intrinsic errors in genome annotation. Trends Genet. 2001;17: 429–431.

79. Jones CE, Brown AL, Baumann U. Estimating the annotation error rate of curated GO database sequence annotations. BMC Bioinformatics. 2007;8: 170.

80. Jones JM, Morrell JC, Gould SJ. Identification and characterization of HAOX1, HAOX2, and HAOX3, three human peroxisomal 2-hydroxy acid oxidases. J Biol Chem. 2000;275: 12590–12597.

81. Dellero Y, Mauve C, Boex-Fontvieille E, Flesch V, Jossier M, Tcherkez G, et al. Experimental evidence for a hydride transfer mechanism in plant glycolate oxidase catalysis. J Biol Chem. 2015;290: 1689–1698.

82. Umena Y, Yorita K, Matsuoka T, Kita A, Fukui K, Morimoto Y. The crystal structure of L-lactate oxidase from Aerococcus viridans at 2.1A resolution reveals the mechanism of strict substrate recognition. Biochem Biophys Res Commun. 2006;350: 249–256.

83. Hackenberg C, Kern R, Hüge J, Stal LJ, Tsuji Y, Kopka J, et al. Cyanobacterial lactate oxidases serve as essential partners in N2 fixation and evolved into photorespiratory glycolate oxidases in plants. Plant Cell. 2011;23: 2978–2990.

84. Rassaei L, Olthuis W, Tsujimura S, Sudhölter EJR, van den Berg A. Lactate biosensors: current status and outlook. Anal Bioanal Chem. 2014;406: 123–137.

85. Richardson EJ, Watson M. The automatic annotation of bacterial genomes. Brief Bioinform. 2013;14: 1–12.

86. Danchin A, Ouzounis C, Tokuyasu T, Zucker J-D. No wisdom in the crowd: genome annotation in the era of big data - current status and future prospects. Microb Biotechnol. 2018;11: 588–605.

87. Knorr S, Sinn M, Galetskiy D, Williams RM, Wang C, Müller N, et al. Widespread bacterial lysine degradation proceeding via glutarate and L-2-hydroxyglutarate. Nat Commun. 2018;9: 5071.

88. Koga Y, Konishi K, Kobayashi A, Kanaya S, Takano K. Anaerobic glycerol-3-phosphate dehydrogenase complex from hyperthermophilic archaeon Thermococcus kodakarensis KOD1. J Biosci Bioeng. 2019;127: 679–685.

89. Weghoff MC, Bertsch J, Müller V. A novel mode of lactate metabolism in strictly anaerobic bacteria. Environ Microbiol. 2015;17: 670–677.

90. Guo X, Zhang M, Cao M, Zhang W, Kang Z, Xu P, et al. d-2-Hydroxyglutarate dehydrogenase plays a dual role in l-serine biosynthesis and d-malate utilization in the bacterium Pseudomonas stutzeri. J Biol Chem. 2018;293: 15513–15523.

91. Gerlt JA. The Need for Manuscripts To Include Database Identifiers for Proteins. Biochemistry. 2018;57: 4239–4240.

92. Griesemer M, Kimbrel JA, Zhou CE, Navid A, D'haeseleer P. Combining multiple functional annotation tools increases coverage of metabolic annotation. BMC Genomics. 2018;19: 948.

93. Erb TJ. Back to the future: Why we need enzymology to build a synthetic metabolism of the future. Beilstein J Org Chem. 2019;15: 551–557.

94. Mazurenko S, Prokop Z, Damborsky J. Machine Learning in Enzyme Engineering. ACS Catal. 2020;10: 1210–1223.

95. Li G, Rabe KS, Nielsen J, Engqvist MKM. Machine Learning Applied to Predicting Microorganism Growth Temperatures and Enzyme Catalytic Optima. ACS Synth Biol. 2019;8: 1411–1420.

96. Wood V, Lock A, Harris MA, Rutherford K, Bähler J, Oliver SG. Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? Open Biol. 2019;9: 180241.

97. Ghatak S, King ZA, Sastry A, Palsson BO. The y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function. Nucleic Acids Res. 2019;47: 2446–2454.

98. Shearer AG, Altman T, Rhee CD. Finding sequences for over 270 orphan enzymes. PLoS One. 2014;9: e97250.

99. Tawfik DS, van der Donk WA. Editorial overview: Biocatalysis and Biotransformation: Esoteric, Niche Enzymology. Curr Opin Chem Biol. 2016;31: v–vii.

100. Qiu H, Geng A, Zhu D, Le Y, Wu J, Chow N, et al. Purification and characterization of a

hemocyanin (Hemo1) with potential lignin-modification activities from the wood-feeding termite, Coptotermes formosanus Shiraki. Appl Biochem Biotechnol. 2015;175: 687–697.

101. Pieslinger AM, Hoepflinger MC, Tenhaken R. Cloning of Glucuronokinase from Arabidopsis thaliana, the last missing enzyme of the myo-inositol oxygenase pathway to nucleotide sugars. J Biol Chem. 2010;285: 2902–2910.

102. Berini F, Casciello C, Marcone GL, Marinelli F. Metagenomics: novel enzymes from non-culturable microbes. FEMS Microbiol Lett. 2017;364. doi:10.1093/femsle/fnx211

103. Robinson SL, Piel J, Sunagawa S. A roadmap for metagenomic enzyme discovery. Nat Prod Rep. 2021 [cited 13 Sep 2021]. doi:10.1039/D1NP00006C

104. Sévin DC, Fuhrer T, Zamboni N, Sauer U. Nontargeted in vitro metabolomics for high-throughput identification of novel enzymes in Escherichia coli. Nat Methods. 2017;14: 187–194.

105. Aravind L. Guilt by association: contextual information in genome analysis. Genome Res. 2000;10: 1074–1077.

106. Jacobson MP, Kalyanaraman C, Zhao S, Tian B. Leveraging structure for enzyme function prediction: methods, opportunities, and challenges. Trends Biochem Sci. 2014;39: 363–371.

107. Christian N, May P, Kempa S, Handorf T, Ebenhöh O. An integrative approach towards completing genome-scale metabolic networks. Mol Biosyst. 2009;5: 1889–1903.

108. Zallot R, Oberg N, Gerlt JA. The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. Biochemistry. 2019;58: 4169–4182.

109. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2020;49: D605–D612.

110. Price MN, Arkin AP. PaperBLAST: Text Mining Papers for Information about Homologs. mSystems. 2017;2. doi:10.1128/mSystems.00039-17

111. Rembeza E, Boverio A, Fraaije MW, Engqvist M. Discovery of two novel oxidases using a high-throughput activity screen. Chembiochem. 2021. doi:10.1002/cbic.202100510

112. Sorokina M, Stam M, Médigue C, Lespinet O, Vallenet D. Profiling the orphan enzymes. Biol Direct. 2014;9: 10.

113. Horiuchi T. Purification and Properties of N-Acyl-d-hexosamine Oxidase from Pseudomonas sp. 15-1. Agric Biol Chem. 1989;53: 361–368.

114. Daniel B, Konrad B, Toplak M, Lahham M, Messenlehner J, Winkler A, et al. The family of berberine bridge enzyme-like enzymes: A treasure-trove of oxidative reactions. Arch Biochem Biophys. 2017;632: 88–103.

115. Hansen OC, Stougaard P. Hexose oxidase from the red alga Chondrus crispus. Purification, molecular cloning, and expression in Pichia pastoris. J Biol Chem. 1997;272: 11581–11587.

116. Omelchenko MV, Galperin MY, Wolf YI, Koonin EV. Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. Biol Direct. 2010;5: 31.

117. Sützl L, Foley G, Gillam EMJ, Bodén M, Haltrich D. The GMC superfamily of oxidoreductases revisited: analysis and evolution of fungal GMC oxidoreductases. Biotechnol Biofuels. 2019;12: 118.

118. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021. doi:10.1038/s41586-021-03819-2

119. Tawfik DS. Enzyme promiscuity and evolution in light of cellular metabolism. FEBS J.

2020;287: 1260–1261.

120. Cheng Q, Sanglard D, Vanhanen S, Liu HT, Bombelli P, Smith A, et al. Candida yeast long chain fatty alcohol oxidase is a c-type haemoprotein and plays an important role in long chain fatty acid metabolism. Biochim Biophys Acta. 2005;1735: 192–203.

121. Savino S, Fraaije MW. The vast repertoire of carbohydrate oxidases: An overview. Biotechnol Adv. 2020; 107634.

122. Cipolatti EP, Cerqueira Pinto MC, Henriques RO, da Silva Pinto JCC, de Castro AM, Freire DMG, et al. Chapter 5 - Enzymes in Green Chemistry: The State of the Art in Chemical Transformations. In: Singh RS, Singhania RR, Pandey A, Larroche C, editors. Advances in Enzyme Technology. Elsevier; 2019. pp. 137–151.

123. Kaushal J, Khatri M, Arya SK. Recent insight into enzymatic degradation of plastics prevalent in the environment: A mini - review. Cleaner Engineering and Technology. 2021;2: 100083.

124. Adams JP, Brown MJB, Diaz-Rodriguez A, Lloyd RC, Roiban G-D. Biocatalysis: A pharma perspective. Adv Synth Catal. 2019. doi:10.1002/adsc.201900424

125. Petkevicius K, Löfstedt C, Borodina I. Insect sex pheromone production in yeasts and plants. Curr Opin Biotechnol. 2020;65: 259–267.

126. Chen C-C, Dai L, Ma L, Guo R-T. Enzymatic degradation of plant biomass and synthetic polymers. Nature Reviews Chemistry. 2020;4: 114–126.

127. Hutchison CA 3rd, Phillips S, Edgell MH, Gillam S, Jahnke P, Smith M. Mutagenesis at a specific position in a DNA sequence. J Biol Chem. 1978;253: 6551–6560.

128. Brannigan JA, Wilkinson AJ. Protein engineering 20 years on. Nat Rev Mol Cell Biol. 2002;3: 964–970.

129. Chen K, Arnold FH. Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. Proc Natl Acad Sci U S A. 1993;90: 5618–5622.

130. Stemmer WP. Rapid evolution of a protein in vitro by DNA shuffling. Nature. 1994;370: 389–391.

131. Arnold FH. Innovation by Evolution: Bringing New Chemistry to Life (Nobel Lecture). Angew Chem Int Ed Engl. 2019;58: 14420–14426.

132. Peisajovich SG, Tawfik DS. Protein engineers turned evolutionists. Nat Methods. 2007;4: 991–994.

133. Shafee T. Evolvability of a viral protease: experimental evolution of catalysis, robustness and specificity. 2 Apr 2014 [cited 25 Oct 2021]. doi:10.17863/CAM.16528

134. Sharma N, Sharma R, Jindal N. Machine Learning and Deep Learning Applications-A Vision. Global Transitions Proceedings. 2021;2: 24–28.

135. Feehan R, Montezano D, Slusky JSG. Machine learning for enzyme engineering, selection and design. Protein Eng Des Sel. 2021;34. doi:10.1093/protein/gzab019

136. Gado JE, Beckham GT, Payne CM. Improving Enzyme Optimum Temperature Prediction with Resampling Strategies and Ensemble Learning. J Chem Inf Model. 2020;60: 4098–4107.

137. Hon J, Borko S, Stourac J, Prokop Z, Zendulka J, Bednar D, et al. EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. Nucleic Acids Res. 2020;48: W104–W109.

138. Hou Q, Kwasigroch JM, Rooman M, Pucci F. SOLart: a structure-based method to predict protein solubility and aggregation. Bioinformatics. 2020;36: 1445–1452.

139. Ryu JY, Kim HU, Lee SY. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc Natl Acad Sci U S A. 2019;116: 13996–14001.

140. Dalkiran A, Rifaioglu AS, Martin MJ, Cetin-Atalay R, Atalay V, Doğan T. ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. BMC Bioinformatics. 2018;19: 334.

141. Wu Z, Johnston KE, Arnold FH, Yang KK. Protein sequence design with deep generative models. Curr Opin Chem Biol. 2021;65: 18–27.

142. Repecka D, Jauniskis V, Karpus L, Rembeza E, Rokaitis I, Zrimec J, et al. Expanding functional protein sequence spaces using generative adversarial networks. Nature Machine Intelligence. 2021;3: 324–333.

143. Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, et al. CATH: increased structural coverage of functional space. Nucleic Acids Res. 2021;49: D266–D273.

144. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Networks. arXiv [stat.ML]. 2014. Available: http://arxiv.org/abs/1406.2661

145. Takahashi-Íñiguez T, Aburto-Rodríguez N, Vilchis-González AL, Flores ME. Function, kinetic properties, crystallization, and regulation of microbial malate dehydrogenase. J Zhejiang Univ Sci B. 2016;17: 247.

146. Rosano GL, Ceccarelli EA. Recombinant protein expression in Escherichia coli: advances and challenges. Front Microbiol. 2014;5: 172.

147. Guo HH, Choe J, Loeb LA. Protein tolerance to random amino acid change. Proc Natl Acad Sci U S A. 2004;101: 9205–9210.

148. Russ WP, Figliuzzi M, Stocker C, Barrat-Charlaix P, Socolich M, Kast P, et al. An evolution-based model for designing chorismate mutase enzymes. Science. 2020;369: 440–445.

149. Hawkins-Hooker A, Depardieu F, Baur S, Couairon G, Chen A, Bikard D. Generating functional protein variants with variational autoencoders. PLoS Comput Biol. 2021;17: e1008736.

150. Wu Z, Kan SBJ, Lewis RD, Wittmann BJ, Arnold FH. Machine learning-assisted directed protein evolution with combinatorial libraries. Proc Natl Acad Sci U S A. 2019;116: 8852–8858.

151. Bedbrook CN, Yang KK, Robinson JE, Mackey ED, Gradinaru V, Arnold FH. Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. Nat Methods. 2019;16: 1176–1184.

152. Ajjolli Nagaraja A, Charton P, Cadet XF, Fontaine N, Delsaut M, Wiltschi B, et al. A Machine Learning Approach for Efficient Selection of Enzyme Concentrations and Its Application for Flux Optimization. Catalysts. 2020;10: 291.

153. Karim AS, Dudley QM, Juminaga A, Yuan Y, Crowe SA, Heggestad JT, et al. In vitro prototyping and rapid optimization of biosynthetic enzymes for cell design. Nat Chem Biol. 2020;16: 912–919.

154. Zhang J, Petersen SD, Radivojevic T, Ramirez A, Pérez-Manríquez A, Abeliuk E, et al. Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism. Nat Commun. 2020;11: 4880.

155. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3: 160018.

156. Swainston N, Baici A, Bakker BM, Cornish-Bowden A, Fitzpatrick PF, Halling P, et al. STRENDA DB: enabling the validation and sharing of enzyme kinetics data. FEBS J. 2018;285: 2193–2204.

157. Bisswanger H. Enzyme assays. Perspectives in Science. 2014;1: 41–55.

158. Lambeth DO, Muhonen WW. High-performance liquid chromatography-based assays of enzyme activities. J Chromatogr B Biomed Appl. 1994;656: 143–157.

159. Bothner B, Chavez R, Wei J, Strupp C, Phung Q, Schneemann A, et al. Monitoring Enzyme Catalysis with Mass Spectrometry*. J Biol Chem. 2000;275: 13455–13459.

160. Acker MG, Auld DS. Considerations for the design and reporting of enzyme assays in high-throughput screening applications. Perspectives in Science. 2014;1: 56–73.

161. Chen KQ, Arnold FH. Enzyme engineering for nonaqueous solvents: random mutagenesis to enhance activity of subtilisin E in polar organic media. Biotechnology . 1991;9: 1073–1077.

162. Bunzel HA, Garrabou X, Pott M, Hilvert D. Speeding up enzyme discovery and engineering with ultrahigh-throughput methods. Curr Opin Struct Biol. 2018;48: 149–156.

163. Longwell CK, Labanieh L, Cochran JR. High-throughput screening technologies for enzyme engineering. Curr Opin Biotechnol. 2017;48: 196–202.

164. Dressler OJ, Casadevall I Solvas X, deMello AJ. Chemical and Biological Dynamics Using Droplet-Based Microfluidics. Annu Rev Anal Chem . 2017;10: 1–24.

165. Gielen F, Hours R, Emond S, Fischlechner M, Schell U, Hollfelder F. Ultrahigh-throughput-directed enzyme evolution by absorbance-activated droplet sorting (AADS). Proc Natl Acad Sci U S A. 2016;113: E7383–E7389.

166. NCATS. Assay Development & Screening. 16 Mar 2015 [cited 28 Sep 2021]. Available: https://ncats.nih.gov/preclinical/drugdev/assay

167. Rembeza E, Engqvist MKM. Adaptation of a Microfluidic qPCR System for Enzyme Kinetic Studies. ACS Omega. 2021;6: 1985–1990.

168. Spurgeon SL, Jones RC, Ramakrishnan R. High throughput gene expression measurement with real time PCR in a microfluidic dynamic array. PLoS One. 2008;3: e1662.

169. Jeske L, Placzek S, Schomburg I, Chang A, Schomburg D. BRENDA in 2019: a European ELIXIR core data resource. Nucleic Acids Res. 2019;47: D542–D549.

170. Hess D, Yang T, Stavrakis S. Droplet-based optofluidic systems for measuring enzyme kinetics. Anal Bioanal Chem. 2020;412: 3265–3283.

171. Hess D, Dockalova V, Kokkonen P, Bednar D, Damborsky J, deMello A, et al. Exploring mechanism of enzyme catalysis by on-chip transient kinetics coupled with global data analysis and molecular modeling. Chem. 2021;7: 1066–1079.

172. Fernandes AC, Gernaey KV, Krühne U. “Connecting worlds - a view on microfluidics for a wider application.” Biotechnol Adv. 2018;36: 1341–1366.

173. Olp MD, Kalous KS, Smith BC. ICEKAT: an interactive online tool for calculating initial rates from continuous enzyme kinetic traces. BMC Bioinformatics. 2020;21: 186.

174. Lo M-C, Aulabaugh A, Jin G, Cowling R, Bard J, Malamas M, et al. Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery. Anal Biochem. 2004;332: 153–159.

175. Home - Olink. 20 Feb 2018 [cited 19 Nov 2021]. Available: https://www.olink.com/

176. Pauling L. The future of enzyme research. Henry Ford Hosp Med Bull. 1956;4: 1–4.

177. Markin CJ, Mokhtari DA, Sunden F, Appel MJ, Akiva E, Longwell SA, et al. Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics. Science. 2021;373. doi:10.1126/science.abf8761

178. Perez-Riverol Y, Bai M, da Veiga Leprevost F, Squizzato S, Park YM, Haug K, et al.

Discovering and linking public omics data sets using the Omics Discovery Index. Nat Biotechnol. 2017;35: 406–409.

179. Russo DP, Zhu H. Accessing the High-Throughput Screening Data Landscape. Methods Mol Biol. 2016;1473: 153–159.

180. Rubiera CO. AlphaFold 2 is here: what's behind the structure prediction miracle. [cited 3 Nov 2021]. Available: https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/